Unsupervised Domain Adaptation for Label-Free Polyp Segmentation in Capsule Endoscopy

Ang Chen Quanzhi Fu Tianshuo Yan Butian Xiong Nuohan Chen Yulin Li
Chinese University of Hong Kong, Shenzhen

Abstract

Capsule endoscopy (CE) is crucial for the diagnosis of gastrointestinal diseases. Due to the limited datasets available for CE and the prohibitive cost of labeling them, there are still relatively few studies that leverage deep learning to analyze CE data. In this project, we propose to use unsupervised domain adaptation (UDA) to achieve label-free polyp segmentation for CE by transferring domain knowledge from an annotated, large-scale traditional endoscopy dataset. We experimented with two types of DA methods: (i) Fourier Domain Adaptation, and (ii) Entropy Minimization, both of which achieved better performance than the unadapted baseline. Our best model, AdvEnt, increased the segmentation Dice score on a customized CE dataset by 6.2%. (Code: `https://github.com/Depersonalizc/vce-patho-discovery`)

1 Introduction

Capsule endoscopy was designed to solve problems of traditional endoscopy, such as gastroscopy and colonoscopy. These endoscopies has a camera connected with a tube to take photos in human body. Although these information are crucial for diagnosis, these methods are invasive and have limited detective range. Capsule endoscopy solved these problems by load a camera into a capsule. The camera could take photos for the whole GI tract of the patients therefore provide more valuable information for gastroenterologists.

The capsule endoscopy also face some challenge. First, examining the recorded video is time consuming. It may take gastroenterologists multiple hours to examine a capsule endoscopy video. Meanwhile, critical information are easy to be missed, as it may only appears in the video for a few seconds. Therefore, researchers are keep develop models to help gastroenterologists increase their efficiency and precision on analyzing capsule endoscopic video. This is also what we try to do.

Polyps is also a hot topic in computer aided diagnosis. Polyps are projecting growth of mucous membrane from surfaces in the body. It is an important indicator to cancerwhile always overlooked in endoscopic analysis. Data shows that about 14%-30% polyps fail to be detected during diagnosis. There are matured model for polyps detection for traditional endoscopy. While research focus on capsule endoscopy are still limited.

The main reason for the small number of studies in this direction is the lack of data. A capsule endoscopy provides a video about 2 hours. And its movement and camera perspective are pretty random, which result in high cost in data annotation. In a VCE dataset called Kvasir-Capsule which contains more than hundreds of thousand images, has only 55 images with polyps labeled. For CE data, only unlabeled images are available. However, the data from traditional endoscopy is sufficient. There are a amount of dataset for polyp segmentation in traditional colonoscopy.

In this project, we use a technique called domain adaptation and developed deep learning models for polyp segmentation in Capsule endoscopy. Assume the traditional endoscopy images come from a source domain and capsule endoscopy images come from a target domain, our models have no

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

requirements on annotated capsule endoscopy data, which provides a data-efficient solution for polyp detection in capsule endoscopy video.

The remaining content of the paper is organized as follows. In section 2, we introduce some related works in semantic segmentation and domain adaptation. In section 3, we demonstrate in details the domain adaptation methods we use. In section 4, we report the experimental results and provide an analysis. In section 5, we conclude our project and propose some future directions of the study.

2 Related Works

2.1 Semantic Segmentation

Semantic segmentation has benefited from the continuous evolution of DNN architectures [15][10]. A segmentation architecture ResUNet++ was proposed in 2019 and achieved state of the art performance in polyp segmentation problem [7]. Semantic Segmentation models usually trained on dataset with pixel-level annotations. Different with classification task where human only need to provide a label to each image, manual annotation is not scalable [16]. Thus, techniques like transfer learning or domain adaptation were used to solve the problem of lack of data.

2.2 Domain Adaptation

Unsupervised Domain Adaptation (UDA) is a well researched topic for multiple tasks like classification and detection. While it was not intensely explored in semantic segmentation until recently.It aims to reduce the gap between two distributions[11].

Adversarial Learning for domain adaptation is one of the most well explored topics. Adversarial learning uses a discriminator trained to maximize the confusion between source and target representations, thus reduce the domain discrepancy [3]. It involves two networks. The first network predict the segmentation maps from a input from source or target domain. Another network is a discriminator. The discriminator takes the feature maps from the segmentation network and tries to preidict domain of the input. The optimization gaol for segmentation network is to deceive the discriminator. Optimize such a problem can making the features from source and target domain have similar distribution [14].

Some methods use a additional generative network to generate target images conditioned on the source. Like Cycle-Consistent Adversarial Domain Adaptation proposed by Hoffman et al [5] and Cycle-GAN [17]. Although these networks shows excellent performance, they usually require more memory and hard to be trained with limited computational resources.

Self-training is also an approach. The idea is to use the prediction from an ensembled model or a previous state of model as pseudo-labels for the unlabeled data to train the current model [14]. It can be used with many semi-supervised methods [9] and can dealing with class imbalance and embedded with spatial prior [18].

3 Methods

In this project, we leverage two promising unsupervised domain adaptation (UDA) methods to tackle the problem of label-free polyp segmentation in capsule endoscopy: (i) a Fourier transformation-based DA technique termed Fourier Domain Adaptation (FDA), (ii) entropy-based DA methods: Direct Entropy Minimization (MinEnt) and Adversarial Entropy Training (AdvEnt).

In the setting of unsupervised domain-adapted semantic segmentation, we are given a source (labeled) dataset $D^s = \{(x_i^s, y_i^s) \sim P(x^s, y^s)\}_{i=1}^{N}$ where $x^s \in \mathbb{R}^{H \times W \times 3}$ is a color image, and $y^s \in \mathbb{R}^{H \times W \times K}$ is the one-hot semantic map associated with x^s. Meanwhile, $D^t = \{x_i^t\}_{i=1}^{N_t}$ denotes the target dataset, where the ground-truth semantic labels are absent. Typically, the segmentation network trained on D^s will experience a significant drop in performance when tested on D^t and domain adaptation algorithms are proposed to reduce the performance drop in D^t.

3.1 Fourier Domain Transformation [15]

Fourier domain adaptation [15] is based on a simple assumption. Data from different domains have only differences in some low-level features: like color, sharpness, contrast, etc. But the high-level semantic information are same. For example, in Figure 1 the left image is taken from a famous video game GTA5. And the right image is a real photo taken in a US city. Although they are largely not same: you can quickly realize that the left image is a screenshot from a video game. But the semantic information are almost same. They are all city landscape, both contain road, vehicles, building, etc. Human being can easily ignore the low-level differences and finding similarities in these two images. However, those low-level differences causes confusion to deep learning model and make its' performance drop dramatically.

Figure 1: left: City scene from a video game GTA5; right: City scene from CityScape: a real city image dataset. They have similar high-level semantic components while different low-level features like contrast, color, etc. (Source: [15])

To shrink the gap generated by different low-level features. Fourier Domain Adaptation utilizing a principle in image processing, which is low-level features are encoded in low-frequency region in frequency domain. With this principle, FDA bridges the domain gap by replacing the low-frequency region of Source by Target. A illustration of the process can be seen in Figure 2. Again, we have a source image from GTA5 and a target image of real cityscape. To transfer the source to the target domain. We first do fourier transformation on both image to get their frequency map. Then we replace source's low frequency region, and apply inverse fourier transformation to obtain the transformed image. As you can see, we finally got a image with source content but have very similar low-level information to target image.

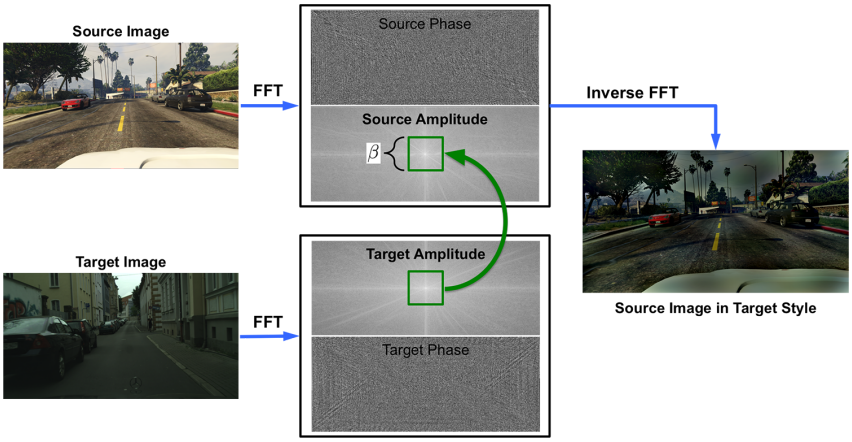

Figure 2: The process of Fourier Domain Adaptation. We apply Fourier transformation to images in both region and replace source image's low-frequency region with the target image's low-frequency region. Then we apply inverse Fourier transformation to obtain the transformed source image. (Source: [15])

The rigorous exchange process can be expressed as follows. Let $\mathscr{F}^A, \mathscr{F}^P : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$ be the amplitude and phase components of the Fourier transform $\mathscr{F}$ of an RGB imgae. We applied

Fourier transform to each channel:

$$\mathscr{F}(x)^{(m,n)} = \sum_{h,w} x^{(h,w)} e^{-j2\pi\left(\frac{h}{H}m + \frac{w}{W}n\right)} \tag{1}$$

where j denotes the imaginary unit.

Given two random sampled images $x^s \sim D^s, x^t \sim D^t$, Fourier Domain Adaptation can be formalized as:

$$x^{s\to t} = \mathscr{F}^{-1}([M_\beta \circ \mathscr{F}^A(x^t) + (1 - M_\beta) \circ \mathscr{F}^A(x^s), \mathscr{F}^P(x^s)]) \tag{2}$$

where $x^{s\to t}$ denotes the transformed image, $\circ$ denotes element-wise multiplication. $M_\beta \in \mathbb{R}^{H\times W\times 3}$ denotes the exchange mask, where elements in index range $[-\beta H : \beta H], [-\beta W : \beta W]$ are 1 and 0 elsewise. $\mathscr{F}^{-1}$ denotes inverse Fourier transform.

In Figure 3 the result of applying Fourier domain adaptation to the Kvasir-SEG source dataset and Video-Kvasir-Capsule target dataset. The training process of FDA is very similar to the traditional

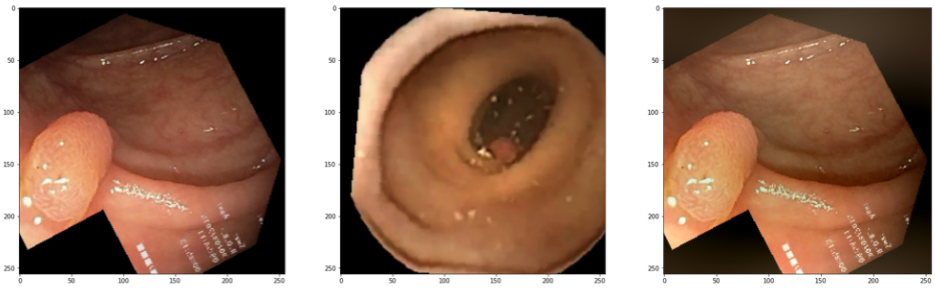

Figure 3: left: source image from Kvasir-SEG dataset; middle: target image from VCE datset; right: transformed source image with Kvasir-SEG content and VCE style.

semantic segmentation algorithm. Cross entropy is selected as loss function.

$$\mathcal{L}(\Phi^\theta; D^{s\to t}) = -\sum_i \langle y_i^s, \log(\Phi^\theta(x_i^{s\to t}))\rangle) \tag{3}$$

where Φ_θ denotes a semantic segmentation network with parameters θ, and $\langle\cdot,\cdot\rangle$ denotes the inner product. The training process can be summarized as follows, in each iteration,

1. Random sampling batch of source and target images X^s, X^t and the segmentation mask of source image Y^s with same batch size from two datasets respectively.
2. Apply equation 2 to X^s and X^t to do the domain transform and got transformed images $X^{s\to t}$.
3. Use $X^{s\to t}$ and Y^s to compute cross entropy loss using formula 3
4. Compute gradients for the network parameters w based on the loss and update the network parameters w.
5. Repeat step 1-4 until convergence

3.2 Entropy-Based Domain Adaptation

Entropy-based domain adaptation aims to bridge the domain gap by enforcing the prediction entropy distribution of the model to be similar between source and target domains. Given a semantic segmentation model Φ_θ with parameters θ, and a sample $x \in \mathbb{R}^{H\times W\times 3}$, we denote by $\hat{y} := \Phi_\theta(x) \in \mathbb{R}^{H\times W\times K}$ the soft segmentation map where each $(\hat{y}^{(h,w)} \in \mathbb{R}^K$ is the outputted class probabilities vector of the pixel $x^{(h,w)}$. (For our purpose of polyp segmentation, $K = 2$.) In information theory, a useful quantity is the self-information $I := -\log \hat{y}$, which is closed connected to the notion of the entropy map, $\mathcal{H}$, by

$$\mathcal{H}^{(h,w)} = \langle \hat{y}^{(h,w)}, I^{(h,w)}\rangle, \qquad \forall h \forall w \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. [14] observes that a model trained solely on the source domain produces over-confident predictions in the source domain and under-confident predictions in the target domain. This corresponds to low-entropy source predictions and high-entropy target predictions (Figure 4).

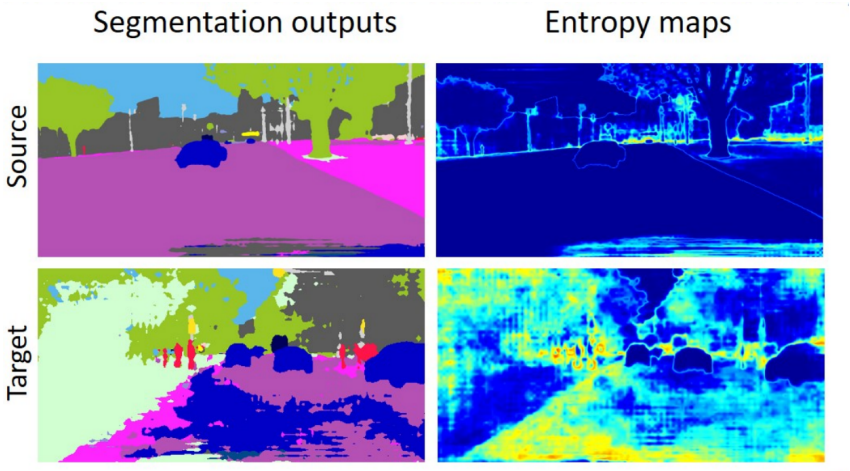

Figure 4: Predicted segmentation maps and entropy maps from the model trained solely in the source domain. (Source: [14])

Inspired by this observation, we apply two entropy minimization techniques to unsupervisedly adapt the model trained on the source domain to the target: (i) Direct Entropy Minimization (MinEnt) and (ii) Adversarial Entropy Training (AdvEnt). We shall examine each in more details in the following sections.

3.2.1 Direct Entropy Minimization

Since the source-trained model produces under-confident predictions on the target domain, MinEnt puts an entropy regularization term to encourage the model to output confident predictions similar to that in the source domain. Specifically, we introduce the entropy loss

$$\mathcal{L}_{\text{Ent}}(x^t) = \frac{1}{HW} \sum_{h,w} \mathcal{H}^{(h,w)}(x^t) \tag{5}$$

to each unlabled samples $x^t \in D^t$. On each labeled samples $(x^s, y^s) \in D^s$, we apply the conventional cross-entropy segmentation loss

$$\mathcal{L}_{\text{Seg}}(x^s, y^s) = \frac{1}{HW} \sum_{h,w} \langle -\log(y^s)^{(h,w)}, (\hat{y}^s)^{(h,w)} \rangle \tag{6}$$

where $\log(\cdot)$ is applied element-wise. During MinEnt training, we jointly optimize the segmentation loss in the source domain and the entropy loss in the target domain, by solving

$$\min_{\theta} \mathop{\mathbb{E}}_{(x^s,y^s)\sim D^s} [\lambda_{\text{Seg}} \mathcal{L}_{\text{Seg}}(x^s, y^s)] + \mathop{\mathbb{E}}_{x^t \sim D^t} [\lambda_{\text{Ent}} \mathcal{L}_{\text{Ent}}(x^t)] \tag{7}$$

where $\lambda_{\text{Ent}}, \lambda_{\text{Seg}}$ are the weighting factors.

3.2.2 Adversarial Entropy Training

Observe from Eq. (4) that the entropy map H can be interpreted as a sum of the self-information I across classes, weighted by the class probabilities. Directly minimizing the entropy thus ignores the internal structure of the class information. It is desirable to have a network automatically learn the most efficient way to combine the predicted self-information. Furthermore, minimizing the aggregated entropy over all pixel locations may not be the best way to match model's predictions in

the source domain. Learning to mimic complex patterns within the entropy map should benefit the performance.

To this end, we use adversarial training that simultaneously optimizes the segmentation model Φ_θ and a discriminator D_ω that tries to distinguish source/target predictions from the model Φ_θ, while Φ_θ attempts to fool D_ω. In this work, the discriminator is implemented as a fully-convolutional network that takes as input a class-information map (that could either be from a source or target prediction)

$$\mathcal{I}^{(h,w)} = \hat{y}^{(h,w)} \circ I^{(h,w)}, \qquad \forall h \forall w \tag{8}$$

and outputs the patch-wise probabilities $D_\omega(\mathcal{I}) \in \mathbb{R}^{H' \times W'}$ of each patch being from a source prediction. Here, $\circ$ is multiplication applied element-wise. Define the likelihood functions:

$$\begin{aligned} \mathcal{L}_s(x^s) &= \frac{1}{H'W'} \sum_{h,w} \log D_\omega(\mathcal{I}(x^s))^{(h,w)} \\ \mathcal{L}_t(x^t) &= \frac{1}{H'W'} \sum_{h,w} \log \left(1 - D_\omega(\mathcal{I}(x^t))^{(h,w)}\right) \end{aligned} \tag{9}$$

During training, we jointly optimize Φ_θ and D_ω by solving the min-max problem

$$\min_\theta \max_\omega \mathop{\mathbb{E}}_{x^t \sim D^t} [\lambda_{\text{Adv}} \mathcal{L}_t(x^t)] + \mathop{\mathbb{E}}_{(x^s, y^s) \sim D^s} [\lambda_{\text{Adv}} \mathcal{L}_s(x^s) + \lambda_{\text{Seg}} \mathcal{L}_{\text{Seg}}(x^s, y^s)] \tag{10}$$

4 Experiments

4.1 Experimental Details

4.1.1 Data and Preprocessing

In this project, we choose Kvasir-SEG, a public traditional endoscopy polyp segmentation dataset, as our source-domain dataset. We manually collect video capsule endoscopy (VCE) dataset as our target-domain dataset.

Kvasir-SEG The Kvasir-SEG dataset [6] contains 1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2. The resolution of the images contained in Kvasir-SEG varies from 332x487 to 1920x1072 pixels. The images and its corresponding masks are stored in two separate folders with the same filename. The image files are encoded using JPEG compression, and online browsing is facilitated. The open-access dataset can be easily downloaded for research and educational purposes.

VCE Dataset Kvasir-Capsule [13] contains images from inside the gastrointestinal (GI) tract via an capsule endoscopy procedure. The images are carefully annotated by medical experts. Here, we only use the polyps class, and three normal classes in distinct location of GI tract. Considering the limited polyp samples (55) in the dataset, we also scrape another 120 CE polyp images from the internet and annotate them manually (only for testing).

Preprocessing We resize all images and masks to 256×256 and applied random rotation, random flips, and random translation during the training phase.

4.1.2 Implementation Details

We test all our domain adaptation methods using an state-of-the-art poly segmentation network called ResUNet++[7]. We train the baseline model on Kvasir-SEG and test the performance on VCE Dataset directly to get the unadapted baseline performance. For FDA, MinEnt, and AdvEnt, we use the baseline model as initialization and then apply the DA technique to train for 10000 iterations.

We use the PyTorch framework to implement our model with the following parameter settings: **baseline**: mini-batch size 16, learning rate $1e-2$; **Fourier DA and MinEnt**: mini-batch size 16, learning rate $1e-2$; **AdvEnt**: mini-batch size 8, learning rate $1e-3$; **MinEnt+AdvEnt**: mini-batch size 8, learning rate $1e-3$.

Models	Acc(%)	Prec(%)	Recall(%)	IoU(%)	Dice(%)
Baseline (w/o DA)	91.1	62.4	77.8	53.0	69.3
FDA	92.0	67.7	72.7	54.0	70.1
MinEnt	**93.0**	**74.6**	69.6	56.3	72.0
AdvEnt	92.9	72.3	73.1	57.1	72.7
MinEnt+AdvEnt	92.6	67.9	**80.3**	**58.2 (+9.8%)**	**73.6 (+6.2%)**

Table 1: Result Comparison for Different Domain Adaptation Models

4.2 Results

4.2.1 Quantitative Results

The experiments results are summarized in table 1. All unsupervised DA methods outperform the unadapted baseline, although improvement is more significant with the entropy minimization approaches. With the combination of direct entropy minimization and adversarial training, we are able to achieve a 9.8% increase in the IoU as well as a 6.2% increase in Dice Score. FDA shows a poorer performance compared with Entropy based methods. A possible reason is the internal complexity of the CE data. Different from cityscape images, factors including camera position or angle may result in a significant change in image style. In our experiments, we also noticed that even a very small β may lead to obvious artifacts in the transformed image. Therefore, FDA may not be the most promising approach when dealing with CE data.

4.2.2 Qualitative Results

Figure 5 shows the segmentation maps outputed by different methods. It is evident that models after domain adaption produces segmentation mask with sharper, more defined polyp boundaries. Compared with entropy-based results, FDA sometimes fail to output a segmentation. This may due to during the training process, some source images are transformed using target image with very different content. Entropy-based methods have a better performance. The improvement is more evident when we look at the entropy map (Figure 6) of the model prediction. The models with entropy based domain adaptation learns to concentrate the entropy only along the boundary of the polyps, which leads to sharper predictions.

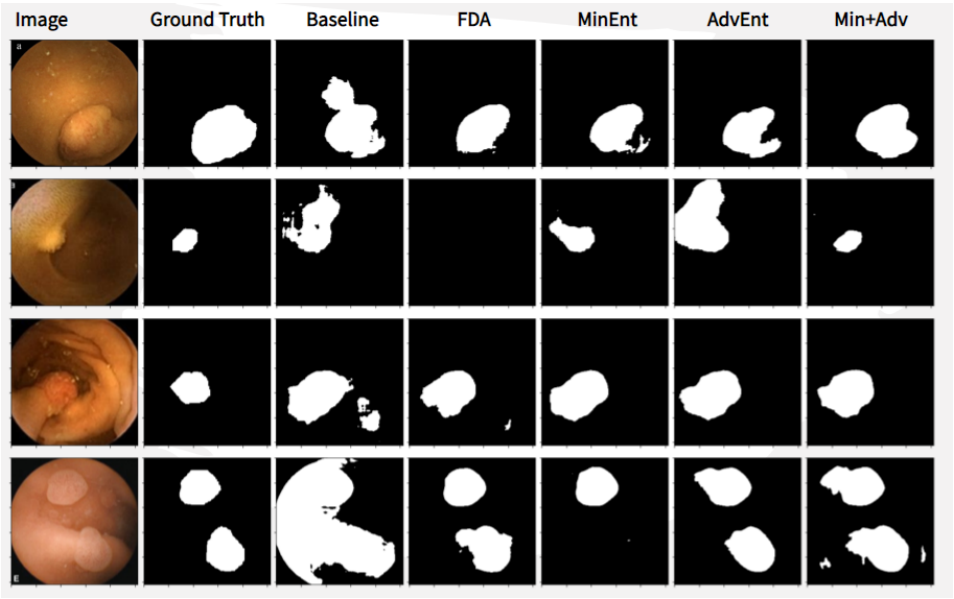

Figure 5: Predicted segmentation maps and entropy maps of the model in the source and target domains.

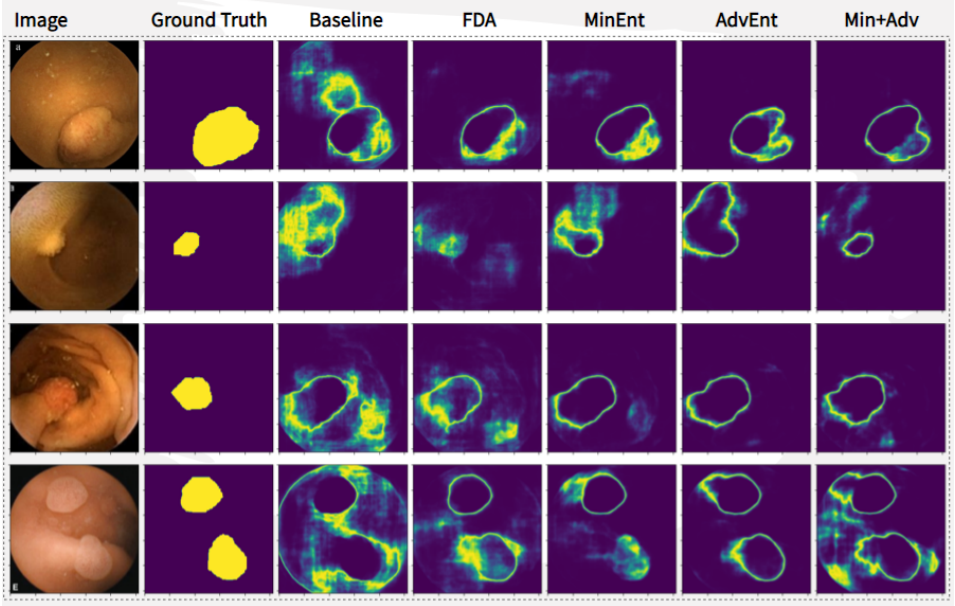

Figure 6: Predicted segmentation maps and entropy maps of the model in the source and target domains.

5 Conclusion & Future Work

In this project, we utilize three domain adaptation methods to tackle the polyp segmentation for capsule endoscopy video. All three domain adaptation shows performance increases compared with the baseline. Among three DA methods, experimental show that the FDA algorithm is not very suitable for the conversion from traditional endoscopy to capsule endoscopy. The two DA methods based on entropy minimization shown a superior performance.

In the future, we plan to collect a larger capsule endoscopy dataset to improve DA performance and validation accuracy. In this work, we mainly focus on inter-domain adaptation, more intra-domain adaptation are also worth to try. Incorporating more label-free learning methods, like self-supervised learning or active learning, is also a direction to further boost the model performance.

A Appendix

We did a lot of experimentation before actually choosing Domain Adaptation as our topic. These attempts include motion deblur in capsule endoscopy, direct localization of polyps, keyframe extraction, and more. Due to data or computing power constraints, these explorations did not end up being our subject. We attach some experimental reports from the exploration process here for future reference.

A.1 Feature Extraction for Important Frame Detection

The detection of the important frames of a video produced by Capsule Endoscopy (CE) is important for the polyps detection, because an extremely large number of frames are generated from the screening procedure of tract (just video recordings, which are able to take up to 30 minutes to 8 hours to complete the process and approximately more than 5 thousand endoscopy frames are produced) in which only a small partition of the frames contain important information – polyps [12]. However, it is a lengthy and hard process for the experts to observe each frame separately, and the abnormal frames can be easily overlooked in the procedure. Therefore, there is a requirement for automatic polyps clinical practitioners of CE-produced frames, which is usually realized by the Computer-aided diagnosing (CAD) systems based on varient machine-vision algorithms.

A.1.1 Feature Extraction Strategies

Although according to researchers [1], there are few researches target on automatic important frames detection of polyps by feature extraction algorithms, the feature extraction strategy could gain advantage on acceptable interpretability compared to other deep learning based algorithms like CNN, LSTM based models. The extracted features theirselves have mathematical or even biological meanings indicated by the algorithms they base on, which might provide insights for the visual properties of the abnormal frames. Moreover, these features later can be used for the segmentation (as what our groups did), classification, and retrieval of the images. In the view of computation itself, the feature extraction reduces the dimension of every matrix-form image frame into linear vector of features, which significantly saves the computational source and accelerates the classification process.

There are various features of colon images, which could be categorized into spatial domain and frequency domain. The frequency domain based algorithms have been implemented enough in the other part of our research, thus here we will mainly focus on the spatial domain features. A digital image is represented by a two dimension matrix in the spatial domain of image processing. The spatial refers to directly manipulation and analysis of the values in the matrix (pixels). So the spacial features relate to the location of the pixels, every pixel is precisely investigated [4]. Those spacial features could provide the internal location structure of the pixels. Since we assume we do not have prior knowledge of each image frame, we will use several commonly used spacial features of images in the proposed research:

- Local Binary Patterns (LBP): representing images'texture, with the basic form calculated by comparing neighboring entries with the central entry and assigns binary code to it. Then these binary codes are transformed into decimal numbers. Forming a histogram, the texture of image is represented locally by occurrence of these codes. An significant advantage is the rotation invariance.
- Shape-Based Features: fractal dimension, smooth spiral curve, Koch snowflake, Sierpinski triangle can be used for shape feature of an image. Higher order local auto-correlation (HLAC) features are also used for some geometrical features. Several geometric features were used to classify the colon polyps [8].
- Color Histograms: hinting the likelihood of a pixel intensity and the distribution of colors.
- Statistical Texture Features: statistical measures are widely used to represent image texture. Energy, contrast, correlation and homogeneity could be calculated from the image frames as the extracted features. The meaning of the features can be seen from the statistical meaning.

In the following part, we will explain how to utilize these extracted features for polyps detection.

A.1.2 Methods and Experiments

Dataset: Kvasir.

Kvasir is a dataset containing images from inside the gastrointestinal (GI) tract via an endoscopy procedure. The images are carefully annotated by medical experts from Vestre Viken Health Trust in Norway and the Cancer Registry of Norway (CRN). It contains several classes showing anatomical landmarks, phatological findings or endoscopic procedures in the GI tract. The dataset consists of images of different resolutions from 720x576 to 1920x1072 pixels, organized in a way that they are sorted in separate folders named according to the content. Here, we only use the polyps class, and three normal classes in distinct location of GI tract. The distinct locations are suitable as the negative label classes, because this matches the real life application scenario. We use the images instead of a video because there is no time domain information used in our method, and the data is annotated more precisely.

Data Processing:

The images go through pre-processing by normalization and contrast enhancement in order to enhance the performance. Then we extract the features from each image in the dataset.

Since the prior knowledge is not known in application scenario, all the features mentioned in above part are selected for targeted extracted features, such as Local Binary Patterns, fractal dimension (indicating geometrical property of the boundary) and so on. All these feature extraction algorithms are relatively mature in implementation.

Table 2: Results

Model	Sensitivity	Specificity	AUC
Kernel SVM	54.5%	84.6%	0.642
Naïve Bayes	58.7%	94.8%	0.660

After the extraction, an image is actually transformed from a 3-D tensor (RGB image) to a linear vector reflecting internal properties. PCA is performed to further reduce the dimension to 4 (achieve the weight of PCs larger than 75%) and extract the important features. For the classification part, the Kernel support vector machine (Kernel-SVM) and naïve Bayes are selected for the binary classification task in order to gain a acceptable explanability. In detail, NuSVC (Nu-Support Vector Classification) in Scikit-learn is chosen for the task, with kernel RBF to realize the non-linear classification. Deep learning algorithms based models are not selected for this task because of their low interpretability. The sensitivity, specificity and AUC are utilized to evaluate the model performance because this is a imbalanced task.

A.1.3 Results and Discussion

The evaluation results of the classification are shown in Table 1. As we can see in Table 1, the sensitivity and AUC of both models are bad, close to fifty percent. These results indicates that the algorithms can not get a acceptable prediction of the positive samples, or in practice, the significant frames in the CE video. The relatively high specificity results from the imbalanced data instead of that the model suits the structure of the negative samples from this point of view. And these models can not be applied to CE-video important frame detection with an acceptable result.

One of the possible reasons is that the data pre-processing procedure is not enough. Frames acquired from endoscopy normally suffer from kinds of noises and variations such as illumination invariance, scale invariance, rotation invariance and so on. Other noises like poor cleansing and bubbles commonly exist, disturbing the analysis. And the normalization and contrast enhancement can not deal with above issues.

Another important reason may be that the extracted features can not generalize to all polyps well. That is, not all polyps share similar mathematical or statistical structures in their image. Then the features extracted naturally show distinct distributions. However, this needs further investigation because there is not enough computational sources in this research. And there exist multiple steps which could be enhanced. Deep learning based models like CNN-based or transformer-based models can be utilized to investigate the effectiveness of extracted features and to discover the internal nature of the mathematical and statistical structure of polyps CE-frames in the future.

A.2 YOLO model

A.2.1 Data Preparation

Our target is to check the performance of YOLO-v5[2] model using 2-fold cross validation of the Kvasir-Capsule dataset. The YOLO-v5 model is the latest version of YOLO model, which is developed by the team of a company called Ultralytics. In this experiment, we used YOLO-v5-v6.1. The data set is a part of the whole Kvasir-Cpasule dataset. This specific set of images is named "Labelled images" with 47284 pictures in total. We merely focus on the Luminal images, consisting of 11 classes: Angiectasia, Blood - fresh, Blood - hematin, Erosion, Erythema, Foreign Body, Lymphangiectasia, Normal clean mucosa, Polyp, Reduced Mucosal View, Ulcer. We first exclude the "Normal clean mucosa" and "Reduced Mucosal View" class, because the normal class without labelling the lesions are useless in YOLO model. Also, because the insufficiency of the image number, "Polyp" and "Blood - hematin" class are removed. The image distribution of the remaining classes is shown in Figure 7.

Another important point of data preparation lays on the dataset partition. For many images are from the same video, some of which are pictures taken in a continuous time slot. This series of images assembles each other and will leads to a over-fitting model in terms of this specific class. This will influence the evaluation of the performance of the YOLO model as well, on account of the existence

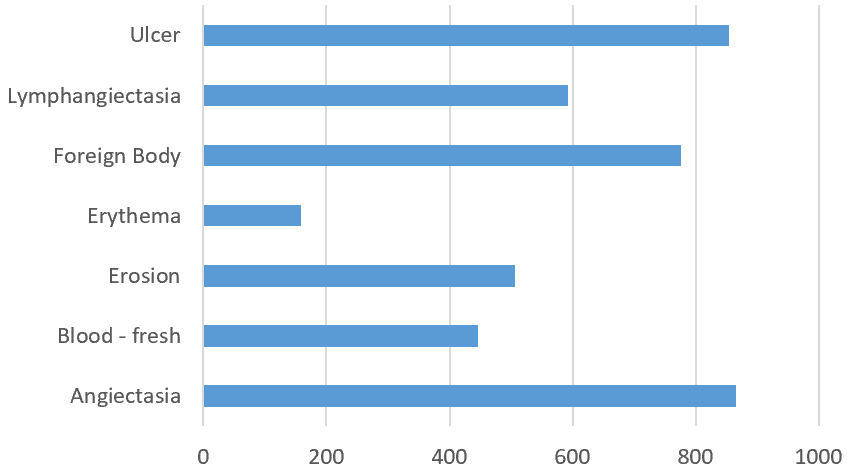

Figure 7: YOLO model classes.

Table 3: The partition of dataset based on vedio

Class	Video numbers	Video of the first validation set
Erosion	9	'0728084c8da942d9', 'dac1e27f7e4d4ef5', 'eb0203196e284797', '5e59c7fdb16c4228', '8ebf0e483cac48d6'
Angiectasia	6	'eb0203196e284797', 'd369e4f163df4aba', '64440803f87b4843'
Erythema	3	'5e59c7fdb16c4228'
Blood - fresh	2	'd369e4f163df4aba'
Foreign Body	4	'8885668afb844852', '7a47e8eacea04e64'
Ulcer	3	'7a47e8eacea04e64'
Lymphangiectasia	3	'7ad22d50ebaf4596'

of similar images appears in both training dataset and validation dataset. The following is the table showing how we divide the dataset into 2 parts based on the video code 3.

A.2.2 Model training and validation

The Code for YOLO-v5 is available from is available from:

`https://github.com/ultralytics/yolov5`

We are supposed to have a deblurred image set based the DeblurGAN model, and use the Yolo-v5 model as a criteria to see how this deblurring model performs. However, because the labelled images do not contain blurred images, it is impossible to realize this idea in this dataset.

We trained the model based on the partition introduced. The result are shown in figure 8. Using the weights file, we developed a lesion detecting app. Figure 9 is a sketch of the app.

References

[1] Hussam Ali, Muhammad Sharif, Mussarat Yasmin, Mubashir Husain Rehmani, and Farhan Riaz. A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract. *Artificial Intelligence Review*, 53(4):2635–2707, 2020.

[2] Yiming Fang, Xianxin Guo, Kun Chen, Zhu Zhou, and Qing Ye. Accurate and automated detection of surface knots on sawn timbers using yolo-v5 model. *BioResources*, 16(3), 2021.

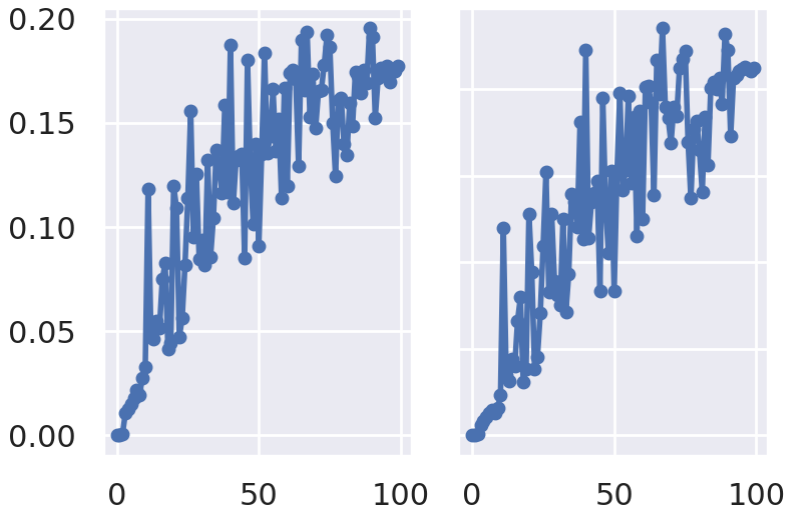

Figure 8: mAP_0.5 for the 2-fold cross validation

Figure 9: YOLO detection on a unlablled image from Kvasir-Capsule dataset

[3] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[4] Kazuhiro Gono, Takashi Obi, Masahiro Yamaguchi, Nagaaki Oyama, Hirohisa Machida, Yasushi Sano, Shigeaki Yoshida, Yasuo Hamamoto, and Takao Endo. Appearance of enhanced tissue features in narrow-band endoscopic imaging. *Journal of biomedical optics*, 9(3):568–577, 2004.

[5] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[6] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[7] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[8] SM Krishnan, Xin Yang, Kap Luk Chan, and PMY Goh. Region labeling of colonoscopic images using fuzzy logic. In *Proceedings of the First Joint BMES/EMBS Conference. 1999*

IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. N, volume 2, pages 1149–vol. IEEE, 1999.

[9] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

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

[11] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[12] Sonu Sainju, Francis M Bui, and Khan A Wahid. Automated bleeding detection in capsule endoscopy videos using statistical features and region growing. *Journal of medical systems*, 38(4):1–11, 2014.

[13] Pia H Smedsrud, Vajira Thambawita, Steven A Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L Eskeland, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):1–10, 2021.

[14] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[15] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[16] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 2020–2030, 2017.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[18] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.