UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Polyp detection and segmentation from endoscopy images

Author:
Mariia KOKSHAIKYNA

Supervisors:
Oles DOBOSEVYCH,
Mariia DOBKO

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Mariia KOKSHAIKYNA, declare that this thesis titled, “Polyp detection and segmentation from endoscopy images” and the work presented in it are my own. I confirm that:

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

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Polyp detection and segmentation from endoscopy images

by Mariia KOKSHAIKYNA

Abstract

Endoscopy is a widely used clinical procedure for the detection of different diseases in internal gastrointestinal tract's organs such as the stomach and colon. Modern endoscopes allow getting high-quality video during the procedure. Computer-assisted methods might support medical specialists in detecting or segmenting anomaly regions on the picture. Many datasets are available and methods to detect polyp regions have been proposed. One kind of task is polyps segmentation on images and videos. The best results in semantic segmentation of polyps are now achieved with fully supervised approaches. In this thesis, we describe experiments with CaraNet model. We checked robustness on cross-validation on several publicly available datasets and small private dataset, tried a few modifications of attention layer in order to improve performance, presented and discussed results.

Acknowledgements

I have been working on this master thesis during my work on the research project on endoscopy image segmentation in UCU ML Lab. I want to thank my supervisors Oles Dobosevych and Mariia Dobko, who helped me during the whole research, and Yurii Yelisieiev, a teammate in this research project who was always open for questions and ready to support.

I also would like to thank Mykhaylo Zayats, a research scientist from IBM Research Europe (Dublin), for initiating the project, and providing the possibility to test our approaches on their private dataset and advice.

Finally, I want to thank Ukrainian Catholic University, Faculty of Applied Sciences, and Oleksii Molchanovskyi, Data Science Academic Program Director.

Contents

List of Figures

List of Tables

List of Abbreviations

GI	**G**astro**i**ntestinal
WL	**W**hite **L**ight
NBI	**N**arrow-**b**and **i**maging
CNN	**C**onvolutional **N**eural **N**etwork
MLP	**M**ulti-**L**ayer **P**erceptron
CRF	**C**onditional **R**andom **F**ield
TTA	**T**est **T**ime **A**ugmentation
MSRF-Net	**M**ulti-**S**cale **R**esidual **F**usion **Net**work
DSDF	**D**ual-**S**cale **D**ense **F**usion
RFB	**R**eceptive **F**ield **B**lock
CFP	**C**hannel-wise **F**eature **P**yramid
PD	**P**artial **D**ecoder
ARA	**A**xial **R**everse **A**ttention
MedT	**Med**ical **T**ransformer
LoGo	**Lo**cal-**G**l**o**bal
JSON	**J**ava**S**cript **O**bject **N**otation
t-SNE	**t**-distributed **S**tochastic **N**eighbor **E**mbedding
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
IoU	**I**ntersection **o**ver **U**nion
FPR	**F**alse **P**ositive **R**ate
FNR	**F**alse **N**egative **R**ate
BCE	**B**inary **C**ross **E**ntropy

Chapter 1

Introduction

1.1 Context

Polyp is an unusual growth of tissue within the inner lining of the stomach or colon. They occur in adult men and women of all ages and usually do not cause symptoms. Most stomach polyps are not cancerous, but there are some which have higher chance of malignancy. Polyps are usually diagnosed during an endoscopy. Endoscope is either a flexible tube with a camera on the end or a capsule camera. Although most stomach polyps do not lead to cancer, some types of polyps need further investigation. In this case, biopsy (tissue sample) can be taken.

To help doctors detect polyps, and further cancer, computer-assisted tools can be used. Creating automated, precise, and robust medical image segmentation methods have been one of the main problems in medical imaging. It is crucial component for computer-aided diagnosis and image-guided surgery systems. Segmentation of organs, lesions, or anomalies from a medical image helps doctors make a correct diagnosis, plan the surgery, and propose treatment plans.

In our work, we research abilities of existing supervised methods for polyps detection and segmentation on different datasets and study how these results could be improved. For now, we do not approach real-time video segmentation solution. Still, polyps segmentation from endoscopy images can be used in verification diagnosis or creation of training materials for doctors.

1.2 Goals of the master thesis

1. To make an overview of the existing state-of-the-art approaches on polyps segmentation and select one of them as a baseline.

2. To explore it's robustness on several public and custom datasets.

3. To formulate hypotheses what modifications can be done to selected approach architecture and conduct experiments.

4. To make conclusions and present results.

1.3 Structure of the thesis

Chapter 2. Medical background This chapter contains general information about colonoscopy, colorectal cancer, and polyps.

Chapter 3. Related works In this chapter we briefly overview existing approaches for medical image segmentation, from classical to supervised. Several state-of-the-art supervised approaches for polyp segmentation which show the best results on HyperKvasir segmented images dataset benchmark are explored in detail.

Chapter 4. Data This chapter contains description of four public datasets on polyp segmentation - HyperKvasir segmented images, CVC-ClinicDB, CVC-EndoSceneStill, and ETIS-LaribPolypDB. We also mention small private custom dataset we run model's inference on. We describe preprocessing algorithm and visualize datasets comparison. We also describe data from EndoCV challenge 2022 we participated in during the project and its preprocessing for our experiments.

Chapter 5. Experiments In this chapter we list all runned cross dataset experiments and experiments with CaraNet architecture and discuss each experiment group results. Experiments from EndoCV challenge 2022 and brief explanation of the results are also added here.

Chapter 6. Conclusions and discussion In this section, we make final conclusions on achieved results and also mention hypotheses for future work.

Chapter 2

Medical background

Polyps The human gastrointestinal (GI) tract consists of different sections, one of them being the large bowel. One of the severe diseases that can affect the large bowel is colorectal cancer. Colorectal cancer is the second most common cancer type among women and the third most common among men[1].

Colorectal cancer usually begins as a polyp, a noncancerous growth that may develop on the inner wall of the colon or rectum as people get older. If not treated or removed, a polyp can become a potentially life-threatening cancer. There are several forms of polyps. Adenomatous polyps, or adenomas, are growths that may become cancerous. They can be found with a colonoscopy. About 10% of colon polyps are flat and hard to find with a colonoscopy unless a dye is used to highlight them. These flat polyps have a high risk of becoming cancerous. Hyperplastic polyps may also develop in the colon and rectum. They are not considered precancerous.[2]

Colorectal cancer Colorectal cancer can begin in either the colon or the rectum. Colorectal cancer begins when healthy cells in the lining of the colon or rectum change and grow out of control, forming a mass called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other body parts. It usually takes years to develop. Both genetic and environmental factors can cause the changes.[3]

Colonoscopy Polyps are predecessors to colorectal cancer. They are found in about half of the patients after 50 years who have a colonoscopy and are increasing with age. Colonoscopy is the procedure for detecting and assessing these polyps with biopsy and removing the polyps. Early disease detection has a significant impact on survival from colorectal cancer. That is why polyp detection is essential. In addition, some studies have shown that polyps are often missed during colonoscopies, with polyp miss rates of 14%-30% depending on the type and size of the polyps. It may depend on endoscopists' skills, as colonoscopy is an operator-dependent procedure. The most frequently missed polyps are flat and smaller polyps (Heresbach et al., 2008). So increasing the detection of polyps decreases the risk of colorectal cancer.

Statistics on colorectal cancer shows that if the cancer is diagnosed at a localized stage, the survival rate is 91%. If the cancer has spread to surrounding tissues or organs and/or the regional lymph nodes, the 5-year survival rate is 72%. If colon cancer has spread to distant parts of the body, the 5-year survival rate is 14%.

Thus, automatic detection of more polyps at an early stage can play a crucial role in improving both prevention of and survival from colorectal cancer. This is the primary motivation behind works on automatic polyp detection and segmentation.

[1] https://www.cancer.net/cancer-types/colorectal-cancer/statistics

[2] https://www.cancer.net/cancer-types/colorectal-cancer/introduction

[3] https://www.cancer.net/cancer-types/colorectal-cancer/diagnosis

Chapter 3

Related work

3.1 General methods overview

Most classical approaches to polyp segmentation are based on contours detection, pixel intensity estimation, filtering and postprocessing (Salman, Ghafour, and Hadi, 2015; Ghosh et al., 2011). Some, use feature descriptors such as geometric and texture descriptors (Figueiredo et al., 2019) while others apply iterated graph cuts (Rother, Kolmogorov, and Blake, 2004). However, all these methods require handcrafting of various parameters per each dataset. Since we aim to develop a generalizable system applicable to any set of new endoscopy data, we decided to avoid classical approaches and move to learnable methods.

State of the art approaches to unsupervised segmentation (Hwang et al., 2019; Kim, Kanezaki, and Masayuki, 2020; Gansbeke et al., 2021) are developed for natural-world images such as in ImageNet, Pascal datasets. One of the risks of using such methods on medical tasks is their adaptivity to binary problems. It is unclear if their performance could be transferred to an anomaly detection pipeline, where only one class (anomaly) is segmented.

Semi-supervised methods depend on generative models which are trained for reconstruction task on general (i.e. ImageNet dataset) set of images. Such trained model can be used for anomaly detection by calculating the deviation between the input with potential anomaly (polyp) and generated output, which tried to reconstruct the image according to learned distribution of normal (healthy) samples. One of the variants of such methods is GANomaly (Akcay, Abarghouei, and Breckon, 2018), which uses a conditional generative adversarial network that jointly learns the generation of high-dimensional image space and the inference of latent space. It was already tested on medical datasets and proved to be efficient in solving anomaly detection task. GANomaly is a semi-supervised approach which doesn't require any per-pixel annotations, thus, it is auspicious to use for medical problems and opens an opportunity to use a larger set of data, since per-pixel labelling is omitted. As authors mention in their paper, the main limitation of this model is its computational complexity since it employs a two-stage approach, and remapping the latent vector is expensive. Another drawback lies in the architecture itself, GANomaly is used to predict the probability of a certain video frame being anomalous, meaning it doesn't produce a segmentation mask for a detected polyp. There are modifications of GANomaly, such as Skip-GANomaly (Akçay and P. Breckon, 2019) which build on the original architecture and improve it.

The best results in semantic segmentation of polyps on HyperKvasir Jha et al., 2020 dataset benchmark[1] are now achieved with fully supervised approaches (Lou et al., 2022; Srivastava et al., 2021; Zhang, Liu, and Hu, 2021).

[1] https://paperswithcode.com/sota/medical-image-segmentation-on-kvasir-seg

3.2 Supervised models for polyp segmentation

All illustrations are taken from the corresponding papers.

ResUNet++ + TTA + CRF

In this paper, authors describe how the ResUNet++ architecture can be extended by applying Conditional Random Field (CRF) and Test-Time Augmentation (TTA) to improve its prediction performance on polyps segmentation (Jha et al., 2019, Jha et al., 2021a).

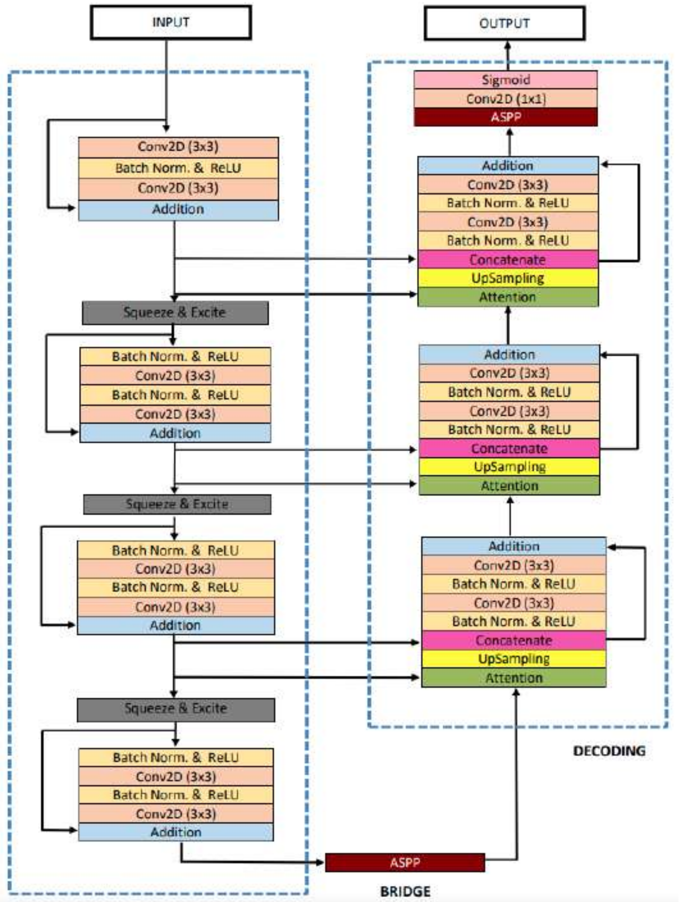

FIGURE 3.1: ResUNet++ architecture

ResUNet++ is a semantic segmentation deep neural network designed for medical image segmentation. The backbone for ResUNet++ architecture is ResUNet. The difference between ResUNet++ and ResUNet is the use of squeeze-and-excitation blocks at the encoder, the ASPP block at bridge and decoder, and the attention block at the decoder - see Figure 3.1.

Authors also introduce a series of additional skip connections from the residual unit of the encoder section to the attention block of the decoder section for the propagation of information. Residual connections improve the training process by directly routing the input information to the output and help to avoid exploding/vanishing gradient during backpropagation. The squeeze and excitation (SE) block is the building block for the CNN that learns the channel weights through global spatial information that increases the sensitivity of the effective feature maps. The feature maps produced by the convolution have only access to the local information, meaning they have no access to the global information left by the local receptive field. To address this limitation, they perform a squeeze operation on the feature maps using the global average pooling to generate a global representation. Then use the global representation and perform sigmoid activation that helps us to learn a non-linear interaction between the channels, and capture the channel-wise dependencies, so the sigmoid activation output acts as a simple gating mechanism. ASPP is used as a bridge between the encoder and the decoder sections, and after the final decoder block to capture the useful multi-scale information between the encoder and the decoder. The attention block gives importance to the subset of the network to highlight the most relevant information. (Jha et al., 2019, Jha et al., 2021a)

Conditional Random Field (CRF) is a statistical modeling method that can model useful geometric characteristics like shape or region connectivity. The use of CRF can improve the models capability to capture contextual information of the polyps and thus improve overall results. In this architecture, they used a dense CRF. (Jha et al., 2019, Jha et al., 2021a)

Test-Time Augmentation (TTA) is a technique of performing reasonable modifications to the test dataset to improve the predictions quality. In TTA, augmentation is applied to each test image, and multiple augmented images are produced. After that, predictions are made on these augmented images, and the average prediction of each augmented image is calculated as the final prediction. In this paper, authors used horizontal and vertical flips for TTA. (Jha et al., 2019, Jha et al., 2021a)

In result, proposed modifications outperform previous ResUNet++ model on the HyperKvasir segmented images dataset and achieve 0.8508 Dice score.

NanoNet

Authors of NanoNet target real-time polyp segmentation in video capsule endoscopy and colonoscopy. It is a lightweight Convolutional Neural Network (CNN) model, computationally efficient and requires less memory. Authors target to create a lightweight model for limited resources constraints for real-time prediction in clinics. Model is optimized for fast inference and high accuracy (Jha et al., 2021b).

The architecture of NanoNet (Figure 3.2) has an encoder-decoder approach. Network uses an pre-trained on ImageNet dataset MobileNetV2 model as an encoder. Authors claim that it helps the model converge faster and achieve higher performance compared to the non-pre-trained model. The decoder is built using a modified version of the residual block. The pre-trained encoder starts with a standard convolution, followed by the bottleneck layer with ReLU6 as the activation function. The entire encoder network progressively downsamples the feature maps by using strided convolution. Decoder uses a bilinear upsampling to increase the spatial dimension (height and width) of the input feature maps. After that, it is concatenated with the appropriate feature maps from the pre-trained encoder using the skip connections. The output of the decoder is fed to a convolution and sigmoid activation (Jha et al., 2021b).

Authors present three NanoNet architectures: NanoNet-A, NanoNet-B, and NanoNet-C. The difference among them is in the number of feature channels. Reduction of trainable parameters simplifies the model complexity and gives a light-weight network (Jha et al., 2021b).

On HyperKvasir segmented images benchmark NanoNet-A model (32, 64 and 128 feature channels) achieves 0.8227 Dice score, authors report (Jha et al., 2021b).

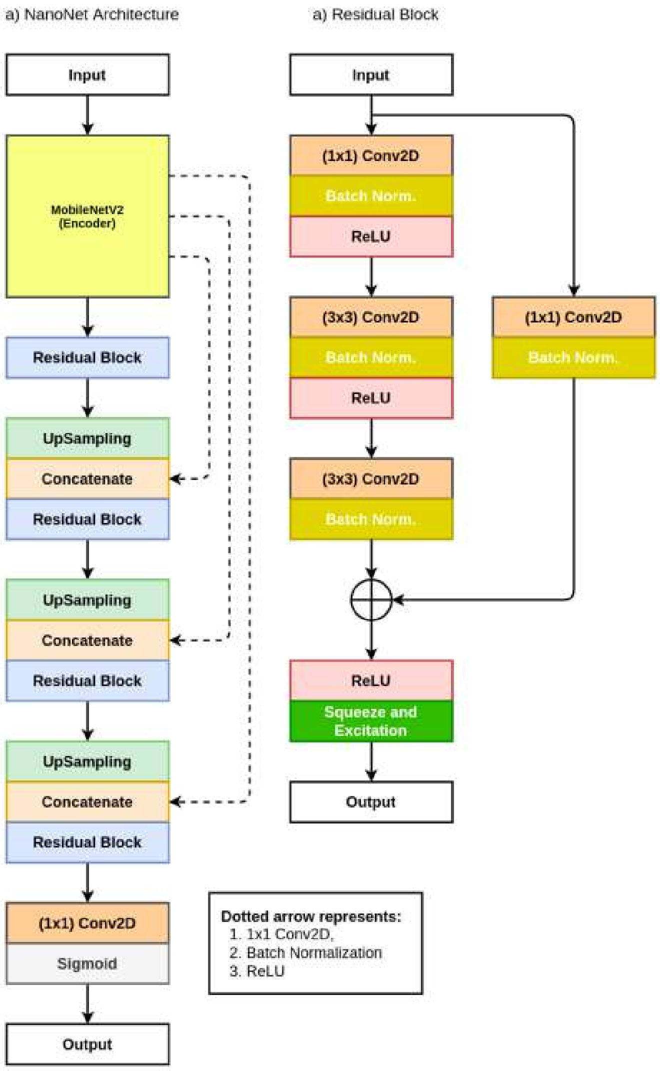

FIGURE 3.2: NanoNet architecture

3.3 State-of-the-art supervised models for polyp segmentation

MSRF-Net

Methods based on convolutional neural networks have shown good performance of biomedical image segmentation. However, most of these methods cannot efficiently segment objects of variable sizes and train on small and biased datasets, which are common for biomedical use cases, claim authors MSRF-Net (Srivastava et al., 2021).

To address the challenges arising with variable sizes, they propose a novel architecture called Multi-Scale Residual Fusion Network (MSRF-Net, Figure 3.3), which is specially designed for medical image segmentation. The proposed MSRF-Net is able to exchange multi-scale features of varying receptive fields using a Dual-Scale Dense Fusion (DSDF) block. DSDF block can exchange information rigorously across two different resolution scales, and MSRF sub-network uses multiple DSDF blocks in sequence to perform multi-scale fusion. This allows the preservation of resolution, improved information flow and propagation of both high- and low-level features to obtain accurate segmentation maps. The proposed MSRF-Net allows to capture object variabilities and provides improved results on different biomedical datasets (Srivastava et al., 2021).

MSRF-Net achieves the dice coefficient of 0.9217, 0.9420, and 0.9224, 0.8824 on Kvasir-SEG, CVC-ClinicDB, 2018 Data Science Bowl dataset, and ISIC-2018 skin lesion segmentation challenge dataset respectively. Although, this approach currently gives a slightly higher result on HyperKvasir segmented images dataset than CaraNet,multi-block structure may lack customization ability. This can make an effect on a generalization ability. And further generalizability tests conducted by authors providea dice coefficient of 0.7921 and 0.7575 on CVC-ClinicDB and Kvasir-SEG, respectively (Srivastava et al., 2021).

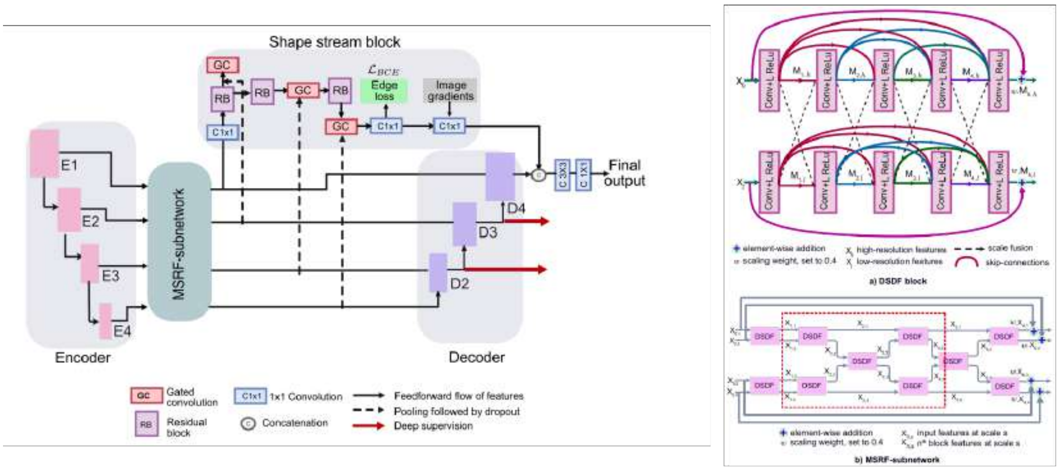

FIGURE 3.3: MSRF-Net architecture

TransFuse

Authors propose a novel parallel-in-branch architecture, TransFuse (Zhang, Liu, and Hu, 2021), to address medical segmentation challenge. TransFuse combines Transformers and CNNs in a parallel style (Figure 3.4), where both global dependency and low-level spatial details can be efficiently captured in a much shallower manner. Besides, a novel fusion technique - BiFusion module is created to efficiently fuse the multi-level features from both branches (Zhang, Liu, and Hu, 2021).

Experiments demonstrate that TransFuse achieves the newest state-of-the-art results on both 2D and 3D medical image sets including polyp, skin lesion, hip, and prostate segmentation, with significant parameter decrease and inference speed improvement. Regarding polyp segmentation, this approach demonstrates 0.918 mean Dice result (same as in CaraNet) on HyperKvasir segmented images dataset (Zhang, Liu, and Hu, 2021).

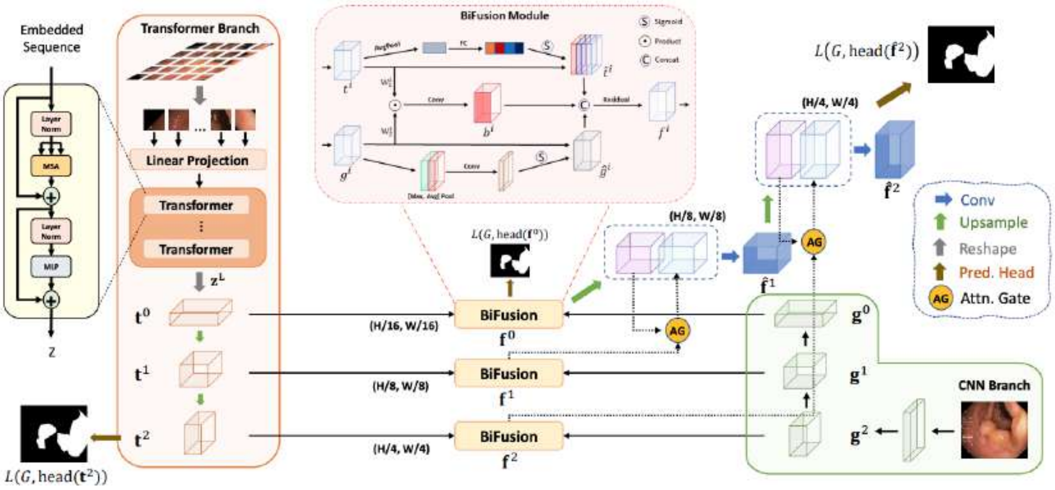

FIGURE 3.4: TransFuse architecture

CaraNet

Authors describe a novel neural network called Context Axial Reserve Attention Network (CaraNet, Lou et al., 2022, Lou and Loew, 2021) to improve the segmentation performance on small objects compared with recent state-of-the-art models. Recently, many CNNs have been designed for segmentation tasks and achieved great success. However most didn't consider the sizes of objects and demonstrate poor performance on small objects segmentation. This can have significant impact on early detection of disease (Lou et al., 2022, Lou and Loew, 2021).

The architecture uses a parallel partial decoder to generate the high-level semantic global map and a set of context and axial reverse attention operations to detect global and local feature information. CaraNet uses Res2Net as a backbone network to extract low- and high-level features. Then it applies a parallel partial decoder to aggregate high-level features. The partial decoder feature is computed by $PD = p_d(f_1, f_2, f_3)$, and they get a global map S_g from the partial decoder. Next, it uses Channel-wise Feature Pyramid (CFP) module to obtain contextual information from high-level features - f_3', f_4', f_5'. Axial reverse attention module contains of two parts: axial attention route and reverse attention route. To capture structural details on tissue, not only approximate location, they erasure foreground object by applying reverse attention $R_i = 1 - Sigmoid(S_i)$. For another route, axial attention is used to keep the global connection and for efficient computing. And the output of A-RA module is represented as $ARA_i = AA \bigodot R_i$ where $\bigodot$ is element-wise multiplication and the AA_i is feature from the axial attention route (Lou et al., 2022, Lou and Loew, 2021).

Loss contains of weighted intersection over union (IoU) and weighted binary cross entropy (BCE) loss components. $\mathcal{L}_{total} = \mathcal{L}(G, S_g^{up}) + \sum_{i=3}^{5} \mathcal{L}(G, S_i^{up})$, where G

is ground-thruth map, and S_i and global map S_g are upsampled to size of G.

The method achieves achieves the top-rank mean Dice segmentation accuracy and shows a distinct advantage in segmentation of small medical objects.

Five polyp datasets: ETIS-LaribPolypDB, CVC-ClinicDB, CVC-ColonDB, CVC-EndoSceneStill and Kvasir are used for evaluation. Authors compare CaraNet with six SOTA medical image segmentation methods: UNet, U-Net++, ResUNetmod, ResUNet++, SFA and PraNet. 80% images from Kvasir and CVC-ClinicDB are used in training set and the remained as a testing dataset. Authors claim that CaraNet not only outperforms compared models according to overall performance, but also on samples with small size polyps. Object's size is considered as a the ratio (proportion) of the number of pixels in the object to whole image.

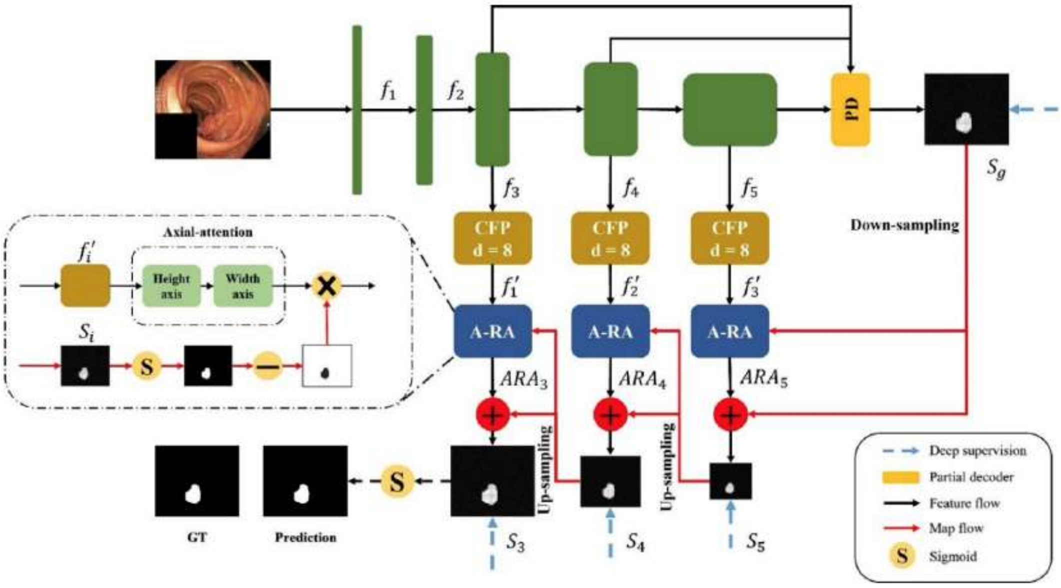

FIGURE 3.5: CaraNet architecture

3.4 State-of-the-art supervised segmentation models not tested for polyp segmentation

Medical Transformer: Gated Axial-Attention for Medical Image Segmentation

In this new paper (Valanarasu et al., 2021), authors explored the use of transformer-based architectures for medical image segmentation. In the result, they proposed a gated axial attention layer which is used as the building block for multi-head attention models - Figure 3.6. On (a) the main architecture diagram of MedT which uses LoGo strategy for training is presented. In (b) the gated axial transformer layer which is used in MedT is shown. And in (c) the Gated Axial Attention layer structure presented, which is the basic building block of both height and width gated multi-head attention blocks found in the gated axial transformer layer (Valanarasu et al., 2021).

They also proposed a LoGo training strategy to train the image in both full resolution as well in patches. The global branch helps learn global context features by modeling long-range dependencies, where as the local branch focus on finer features by operating on patches. Combined, they propose MedT (Medical Transformer) architecture which has gated axial attention as its main building block for the encoder and uses LoGo strategy for training. Unlike other transformer-based model the proposed method does not require pre-training on large-scale datasets. (Valanarasu et al., 2021)

Although authors conducted experiments on three medical datasets (Brain US, GlaS, and MoNuSeg) and achieved a good performance, they didn't try their approach on HyperKvasir segmented images dataset yet.

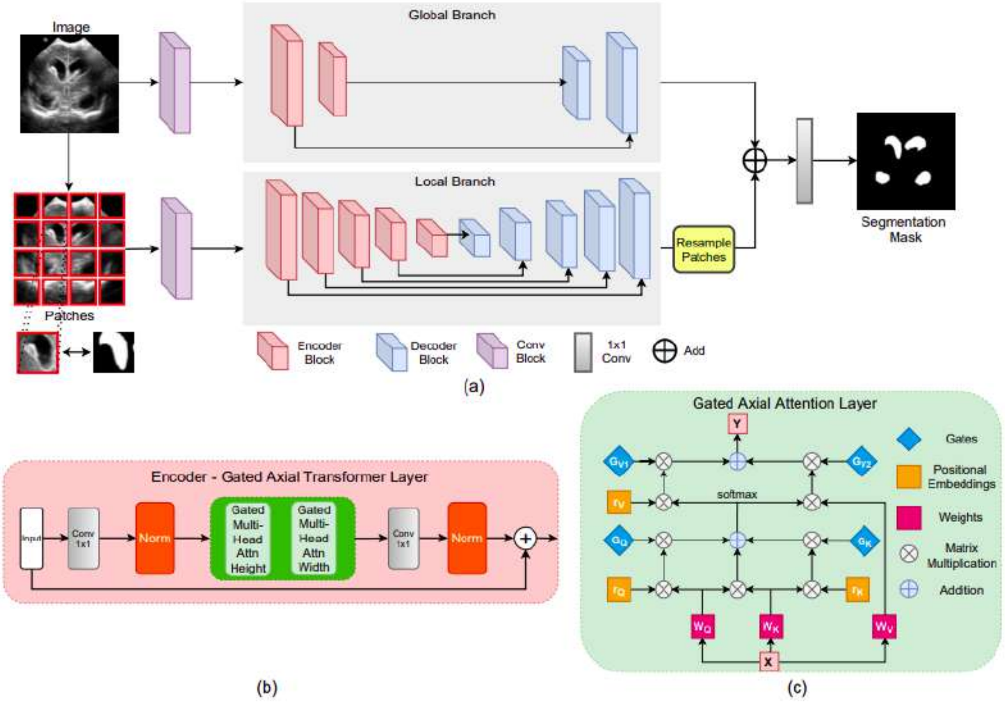

FIGURE 3.6: Medical Transformer architecture

MaskFormer: mask classification-based segmentation

MaskFormer approaches the problem of semantic segmentation as a classification of masks. This approach is an alternative to the per-pixel classification, which predominates in semantic segmentation problems. Instead of classifying each pixel separately, mask classification approaches disjoins the process of semantic segmentation into a division of the image into regions and classification of these regions. MaskFormer is divided into three modules: pixel-level, transformer, and segmentation - Figure 3.7 (Cheng et al., 2021b).

Pixel-level module is an encoder-decoder architecture typically used for the semantic segmentation task. The encoder part (a backbone) generates a high-level feature representation of the image. Further, pixel-level embeddings are obtained by iteratively upsampling feature representation from the encoder (Cheng et al., 2021b).

Transformer module generates N learnable positional embeddings (i.e., queries). This module architecture is adapted from transformers (Vaswani et al., 2017), popular for sequence data. The attention mechanism encodes information about the relation of these segments and enhances them with the image context (Cheng et al., 2021b).

The segmentation module utilizes a linear classifier and a softmax activation function to acquire class probabilities from each query. An MLP (Multi-Layer Perceptron) with two hidden dimensions converts queries into mask embeddings for further conversion. The dot product between mask embeddings and per-pixel embeddings is used to calculate mask predictions (Cheng et al., 2021b).

Model training given matching is performed by utilizing mask classification composed of cross-entropy classification loss and a binary mask loss. Where mask loss is a linear combination of dice and cross-entropy loss (Cheng et al., 2021b).

Although MaskFormer approach has not been tested on medical data, authors report results for ADE20K dataset. The ADE20K semantic segmentation dataset contains more than 20K scene-centric images exhaustively annotated with pixel-level objects and object parts labels. There are 150 semantic categories in total, i.e. sky, road, grass, person, car (Zhou et al., 2017, Zhou et al., 2019). Maskformer achieves a new state-of-the-art of 55.6 mIoU, which is 2.1 mIoU better than the prior state-of-the-art Swin Transformer (Liu et al., 2021). MaskFormer approach also has fewer parameters and faster inference time (Cheng et al., 2021b).

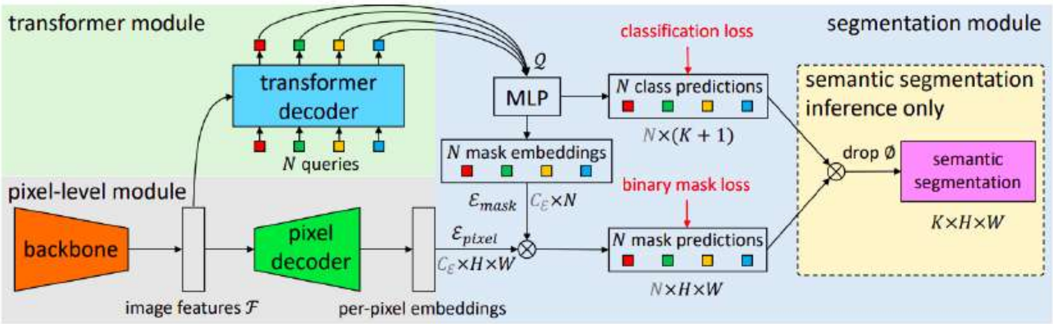

FIGURE 3.7: MaskFormer architecture

Chapter 4

Data

4.1 Public datasets

4.1.1 HyperKvasir segmented images

HyperKvasir dataset (Jha et al., 2020; Borgli et al., 2020), which is the largest multi-class image and video dataset from the gastrointestinal tract available today. The data is collected during real gastro- and colonoscopy examinations at a Hospital in Norway and manually annotated and verified by an experienced gastroenterologist. Hyperkvasir segmented images was the first endoscopic dataset where segmentation masks were created in addition to framewise annotations. This allowed computer vision researchers to contribute in the field of polyp segmentation and automatic analysis of colonoscopy videos.

HyperKvasir dataset's segmented images sectiong provides the original image, a segmentation mask and a bounding box for 1,000 images from the polyp class. In the mask, the pixels depicting polyp tissue, the region of interest, are represented by the foreground (white mask), while the background (in black) does not contain polyp pixels. The bounding box is defined as the outermost pixels of the found polyp. The bounding boxes for the corresponding images are stored in a JavaScript Object Notation (JSON) file. The image and its corresponding mask have the same filename (Jha et al., 2020; Borgli et al., 2020).

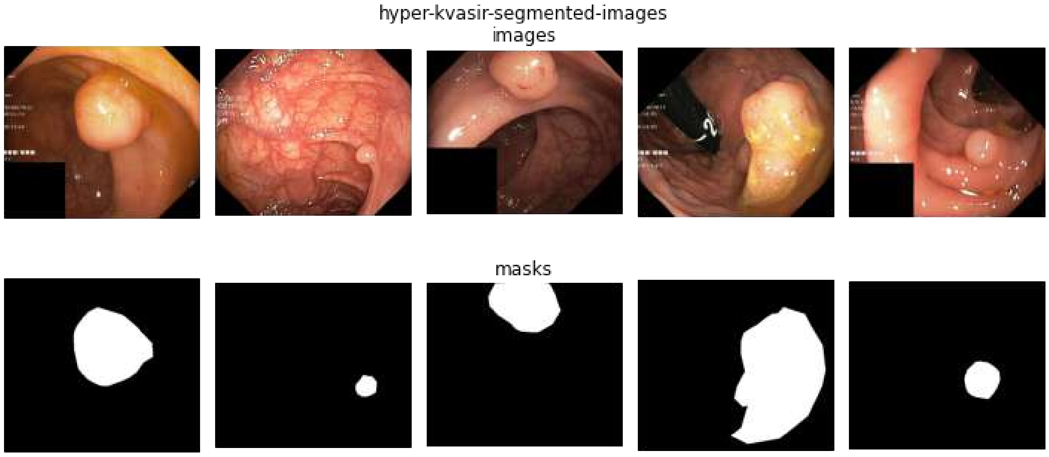

FIGURE 4.1: Examples of images and masks in HyperKvasir dataset

4.1.2 CVC-ClinicDB

CVC-ClinicDB (Bernal et al., 2015; Fernández-Esparrach et al., 2016) is a database of frames extracted from colonoscopy videos. These frames contain several examples of polyps. In addition to the frames, ground truth for the polyps are provided. This ground truth consists of a mask corresponding to the region covered by the polyp in the image.

CVC-ClinicDB is the official database used in the training stages of MICCAI 2015 Sub-Challenge on Automatic Polyp Detection Challenge in Colonoscopy Videos. Overall, dataset contains 612 sequential WL images with polyps extracted from 31 sequences with 31 different polyps acquired from 23 patients.

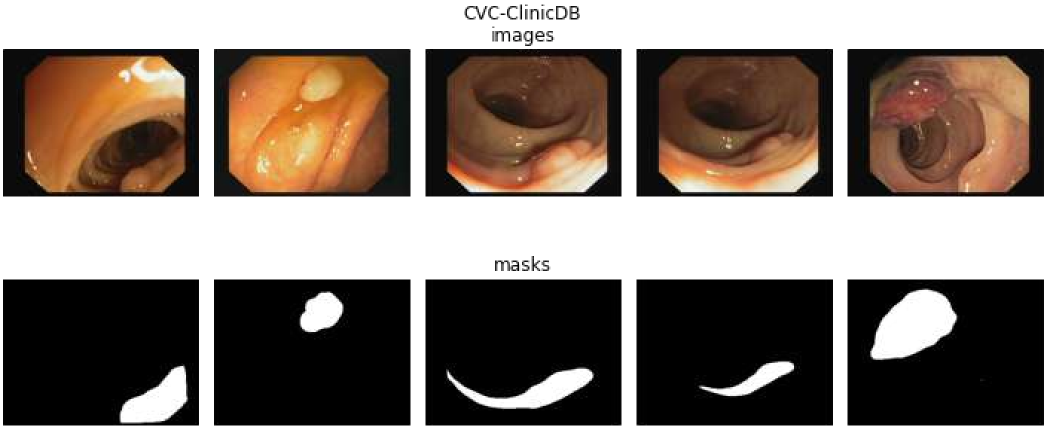

FIGURE 4.2: Examples of images and masks in CVC-ClinicDB dataset

4.1.3 CVC-EndoSceneStill

CVC-EndoSceneStill dataset (Vázquez et al., 2017) is composed of 912 images obtained from 44 video sequences acquired from 36 patients. It has annotations to account for lumen, specular highlights with hand-made pixel-wise annotations, and defined void class for black borders present in each frame. In the annotations, background only contains mucosa (intestinal wall).

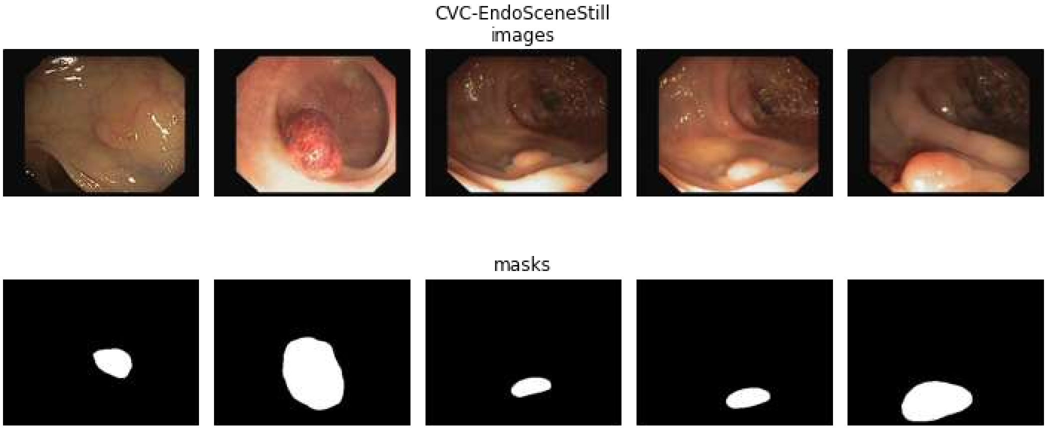

FIGURE 4.3: Examples of images and masks in CVC-EndoSceneStill dataset

4.1.4 ETIS-LaribPolypDB

Dataset contains 196 WL images with polyps extracted from 34 sequences with 44 different polyps. It was created for 'Endovis' Challenge, polyp detection sub-challenge. (Silva et al., 2014)

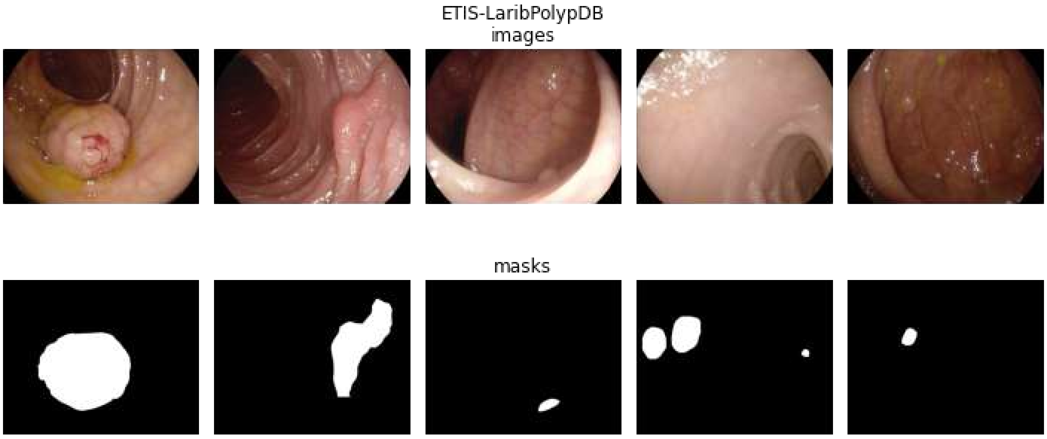

FIGURE 4.4: Examples of images and masks in ETIS-LaribPolypDB dataset

Public datasets information

Statistical dataset information is provided in Table 4.1 (Lou et al., 2022). All datasets images differ in image size and polyp object size.

TABLE 4.1: Public datasets statistics

Dataset	Image size	Number of images	Polyp sizes
HyperKvasir	1070 × 1348	1000	0.79% - 62.13%
CVC-ClinicDB	288 × 384	612	0.34% - 45.88%
CVC-EndoSceneStill	288 × 384, 500 ×574	912	0.30% - 63.15%
ETIS-LaribPolypDB	966 × 1225	196	0.11% - 29.05%

4.2 Custom dataset

We also checked models results on small custom private dataset which is unfortunately not available for us for training due to data usage restrictions. Overall, it contains Stryker Pinpoint endoscopy camera video frames for 20 different patients (1-2 selected frames per patient). For these images there are no segmentation masks available.

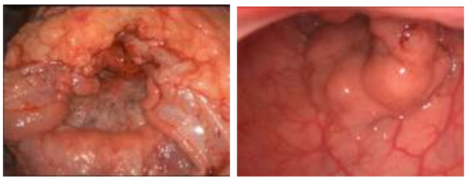

FIGURE 4.5: Examples of images from custom dataset

4.3 Preprocessing

To make training images more similar to custom dataset (no black border, edge rectangles, image of size 480 x 360) we also created a cropped version of each dataset in order to make training camera view more similar to test dataset.

Cropping algorithm is simple:

1. convert image to grayscale
2. mask pixels of intensity less than *threshold_black* (see Table 4.2 for values)
3. mask and remove rows and columns with number of such pixels more than *threshold_black_lines* (see Table 4.2 for values)
4. resize image to target size 480 x 360

As light conditions differ in datasets, we used configurations of thresholds described in Table 4.2.

TABLE 4.2: Configurations of thresholds for cropping.

Dataset	threshold_black	threshold_black_lines
HyperKvasir	15	100
CVC-ClinicDB	20	80
CVC-EndoSceneStill	0.08	90
ETIS-LaribPolypDB	2	180

See examples of original and cropped images from different datasets on Fig. 4.6.

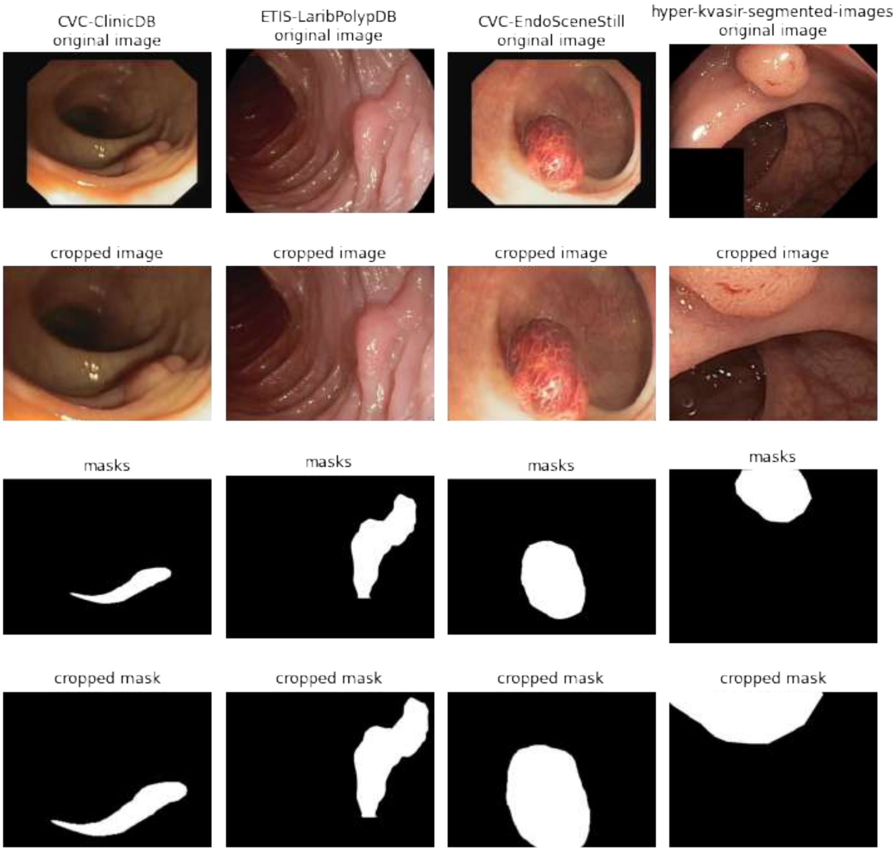

FIGURE 4.6: Examples of cropped images and masks

4.4 Datasets comparison

To visualize differences between images in different datasets we took flattened output of third attention layer of default CaraNet model trained only on HyperKvasir images (both original and cropped versions) and then applied T-SNE (n_components=2, learning_rate='auto', init='random') and UMAP (default configuration) dimensionality reduction algorithms to visualize clustering of image embedding vectors in 2D. Fig. 4.7.

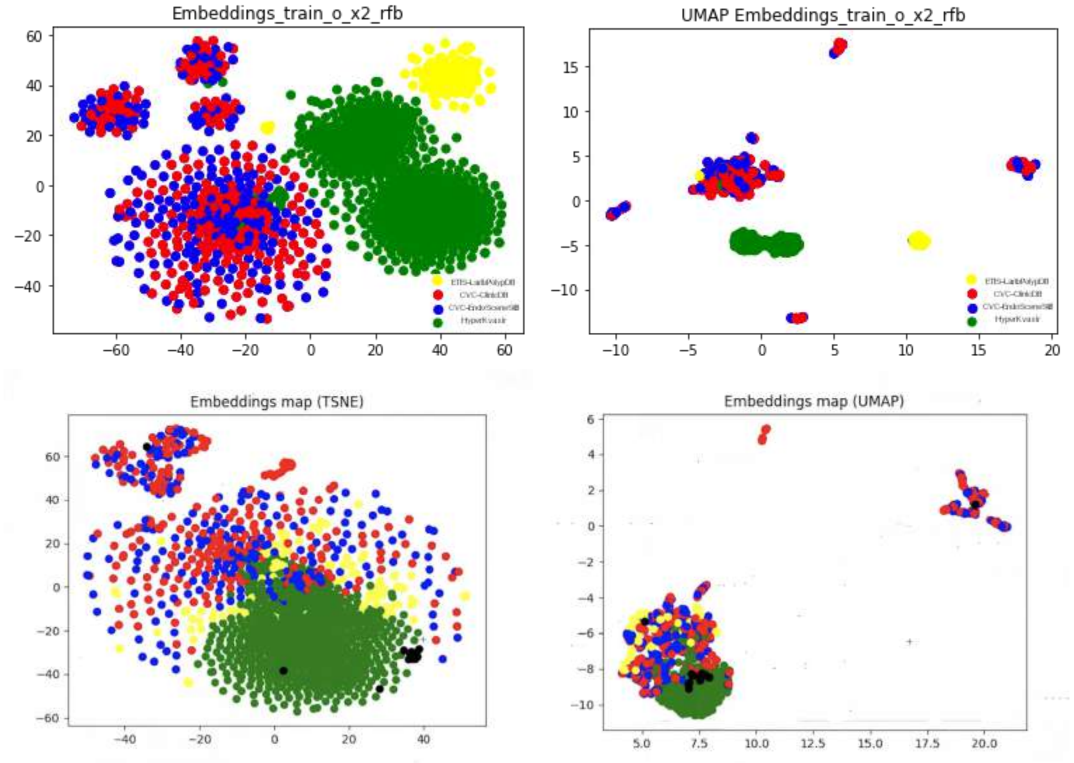

FIGURE 4.7: T-SNE (left column) and UMAP (right columns) image embeddings visualizations on original (first row) and cropped (second row) images. Black points are custom dataset images

From visualizations we can see that original images create strong clusters. And cropped images are more similar between themselves. Representations of custom dataset images are located on the edge of green "HyperKvasir cluster". These difference from "average" image may affect model performance on custom dataset.

Also, we noticed that there are some images in ETIS-Larib dataset, which are more close to HyperKvasir and CVC- images compared to other images in ETIS-Larib. Fig. 4.8.

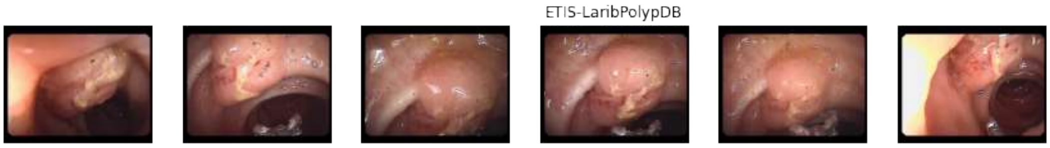

FIGURE 4.8: Nontypical images in ETIS-LaribPolypDB dataset

4.5 EndoCV2022 challenge data

During work on this project, we also participated in 4th International Endoscopy Computer Vision Challenge and Workshop (EndoCV2022). Endoscopic computer vision is a challenge designed to collaborate and curate multicenter datasets and promote building of generalisable models and assess deep learning methods.[1]

This year's challenge consisted of two sub-challenges - (Endoscopy artefact detection) EAD 2.0 and (Polyp generalization) PolypGen 2.0. The data provided is for research purpose only, can be used only for EndoCV2022 challenge participation and will be free to use after the publication of a joint journal paper (Ali et al., 2021a, Ali et al., 2021b, Ali et al., 2021c, Ali et al., 2020).

Proceedings of this challenge are now available online[2].

The PolypGen2.0 subchallenge dataset consists of 46 sequences with 3348 images with polyp labels. Different endoscopes produced these images with various sizes and artifacts - black section located at the left part of the image, blue rectangle with endoscope position, text artifacts, and others (Kokshaikyna, Yelisieiev, and Dobko, 2022). Overall, we can distinguish 15 types of images among these sequences - examples are shown of Figure 4.9. Statistics about different types is shown on Fig 4.10.

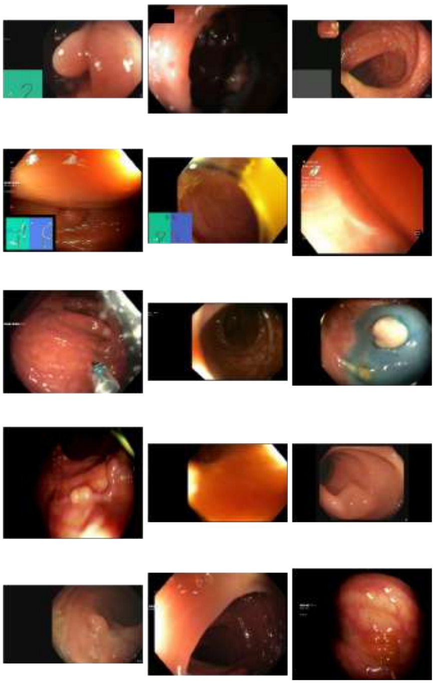

FIGURE 4.9: Examples of different endoscope type images from PolypGen2.0 dataset

For train and validation set, we divided sequences into groups using mannually labeled endoscope image types. For validation set we selected sequences *seq1, seq1_endocv22, seq2_endocv22, seq3, seq3_endocv22, seq5_endocv22, seq7_endocv22, seq10, seq13_endocv22, seq14_endocv22, seq15, seq17, seq19_endocv22, seq21_endocv22*, and *seq24_endocv22*. Other sequences were used in training set. Overall, our train set contained 3306 images and validation set contained 649 images, which is 19,63% of total image number.

[1] https://endocv2022.grand-challenge.org/

[2] http://ceur-ws.org/Vol-3148/

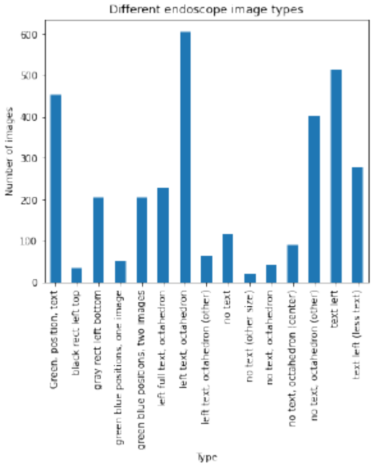

FIGURE 4.10: Different endoscope image types

To bring all images to the same view and use the most informative regions during training, we made simple preprocessing and automatically cropped images cutting black areas on the left and right sides of the input. To do that, we took the center row of the image, sum up values of RGB channels in this row and used a threshold equal to 48. Continuous left and right parts under this threshold were considered redundant and cut. Examples of cropped images are shown in Fig 4.11. This cropping improves the informativeness of images and model generalization.

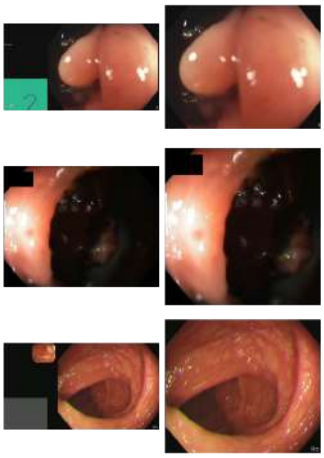

FIGURE 4.11: Examples of cropped images from PolypGen2.0 dataset. First column - before, second - after our pre-processing procedure

Chapter 5

Experiments

Among all existing approaches, we decided to choose CaraNet (Lou et al., 2022), one of the state-of-the-art models for medical image segmentation on HyperKvasir segmented images dataset (Jha et al., 2020), as a main baseline architecture.

We trained it using data from all 4 previously mentioned publicly available for academic purposes endoscopy datasets with polyps, including HyperKvasir (Jha et al., 2020), CVC-ClinicDB, CVC-EndoSceneStill (Bernal et al., 2015; Fernández-Esparrach et al., 2016), and ETIS-LaribPolypDB (Silva et al., 2014).

We examined model's robustness on cross-validation on all datasets, examined a few hypotheses how model architecture changes inspired by other state-of-the-art approaches could affect results. We also checked performance of this model on small custom dataset.

5.1 Scoring methodology

For medical image segmentation task, the most common metric is Dice coefficient. Kvasir-SEG dataset authors also encourage researchers to use it.

$$Dice(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

, where A is predicted mask (set of pixels) and B is ground truth mask.

The Dice coefficient is similar to the IoU, which is also often used in segmentation tasks.

$$IoU(A, B) = \frac{A \cap B}{A \cup B}$$

Both metrics values are from 0 to 1, and a greater value means better performance. In segmentation tasks, Dice is often used as a loss function. Since IoU is not reported for all the models on the Kvasir-SEG leaderboard, we also focused on Dice score more.

Also we measured

$$precision = \frac{TP}{TP + FP}$$

,

$$recall = \frac{TP}{TP + FN}$$

, and

$$FPR = \frac{FP}{N}$$

. FPR is important in medical segmentation because we don't want to miss the positive case, so we target recall. And FPR helps us to control overfitting to recall.

For all experiments, random 60% of images were took for training, 20% for validation and 20% left for testing.

5.2 Cross datasets experiments

We started experiments with default CaraNet architecture, with small updates in training process:

1. added custom *DataLoader* to set up cross validation between datasets "on the fly". It is managed by *–iteration* parameter:
    - 0-3 defines which dataset is used fully for the test fold, others are splitted into train/val parts,
    - -1 to split each of datasets to train/val/test,
    - -100 to use only Hyperkvasir segmented images dataset only.
2. added *ImbalancedDatasetSampler* with possible options:
    - *'dataset'* to sample images from different datasets equally,
    - *'mask_size'* to sample images with small mask proportionally more frequent,
    - and *'only_anomalies'* to skip images with no anomalies after cropping and then sample images from different datasets equally.

Default parameter values are:

- learning rate is 1e-4,
- optimizer is Adam,
- batch size is 6,
- image train size is 352×352.
- gradient clipping margin is 0.5,
- decay rate is 0.1,
- decay epoch is 50,
- augmentations are RandomRotation (up to 90 degrees) and RandomFlip (vertical and horizontal, both with probability 0.5),

Default CaraNet's multi-scale training strategy (*size_rates = [0.75, 1.0, 1.25]*) was used in all experiments, except experiments with gated axial attention module in the next section, where it is stated clearly.

Default CaraNet normalization coefficients were used for all datasets.

Training conditions:

1. one NVIDIA GeForce GTX TITAN X was used for model training,
2. training time took approximately 6-7 hours / 50 epochs,
3. inference time is approximately 0.12 seconds per image (i.e. for 363 images it took 43 seconds).

Cross-dataset experiments In this section of experiments 1-4, we used 3 datasets in training process and 1 in testing. The idea of this was to test model's generalization ablity. From results of experiments, we see noticeable drop in model performance for ETIS-LaribPolypDB dataset. Our assumption was, the reason is smaller average polyp size in this dataset - so we need to focus attention on small polyps in further experiments.

TABLE 5.1: CaraNet experiments 1-4

№	Experiment	Test dataset	Dice on validation	Scores on test
1	CaraNet default + Imbalance Sampler (CV on 4 datasets)	CVC-ClinicDB	0.9038 (CVC-EndoSceneStill 0.9209, ETIS-Larib 0.905, HyperKvasir 0.8931)	Dice 0.9367 precision 0.9283 recall 0.9493
2	–	CVC-EndoSceneStill	0.8943 (ETIS-Larib 0.8976, HyperKvasir 0.8752, CVC-ClinicDB 0.9244)	Dice 0.9399 precision 0.9378 recall 0.9465
3	–	ETIS-Larib	0.9124 (HyperKvasir 0.8901, CVC-EndoSceneStill 0.9305, CVC-ClinicDB 0.9305)	Dice 0.5542 precision 0.5415 recall 0.8242
4	–	HyperKvasir	0.921 (ETIS-Larib 0.914, CVC-EndoSceneStill 0.9221, CVC-ClinicDB 0.9221)	Dice 0.8387 precision 0.8320 recall 0.9119

Experiment on all datasets In experiment 5, we checked that model's training is stable and shows no sings of overtraining - train, validation and test scores doesn't differ significantly.

Experiments with Imbalance Sampler by mask size In experiments 6-7 we added Imbalance Sampler by mask size in order to try to handle model's bad performing on ETIS-LaribPolypDB dataset. Dice score when this dataset is used only in test increased from 0.55 to 0.60 which is rather significant improvement. Imbalance Sampler by mask size also increases overall model performance (experiment 7) slightly.

Inference on custom dataset Although results on public datasets are quite good,

TABLE 5.2: CaraNet experiments 5

№	Experiment	Test dataset	Dice on validation	Scores on test
5	CaraNet default + Imbalance Sampler (each of datasets is splitted into train/val/test)	All	0.9109	Dice 0.9026 precision 0.9086 recall 0.9166

TABLE 5.3: CaraNet experiments 6-7

№	Experiment	Test dataset	Dice on validation	Scores on test
6	CaraNet default + Imbalance Sampler by mask size	ETIS-Larib	0.9049	Dice 0.6087 precision 0.5655 recall 0.8204
7	CaraNet default + Imbalance Sampler by mask size	All	0.9099	Dice 0.9032 precision 0.9131 recall 0.9143

on custom dataset model struggled to find a polyp, and experiment 8 with cropped images which have to be more similar to private ones didn't help either.

While state-of-the-art approaches may show great performance in polyp detection and segmentation when trained and tested on publicly available datasets, there is a large gap between images from real colonoscopy and those in public datasets (Sun et al., 2021). Usually, public datasets contain carefully picked clear images with reasonably sized polyps that stand out from the background, without various artifacts often present in real operations. On the other hand, a considerable portion of the images in colonoscopy operations has different degrees of blurring due to the movement of the camera and intima, camera out-of-focus, or water flushes during an operation. Also, the images frequently contain various artifacts such as fluid, debris, bubbles, reflection, specularity, contrast, saturation, and medical instruments (Ali et al., 2021a, Ali, 2019). Moreover, a wider variety of polyps with different sizes, shapes, or textures can occur in a colonoscopy than in public datasets.

TABLE 5.4: CaraNet experiments 8

№	Experiment	Test dataset	Dice on validation	Scores on test
8	CaraNet default (croppped images) + Imbalance Sampler only anomalies	All	0.8975	Dice 0.8759 precision 0.8440 recall 0.9013 fpr 0.0391

The best and the worst (using Dice metric) segmentation results of experiment 8 can be seen in Figure 5.1 and Figure 5.2.

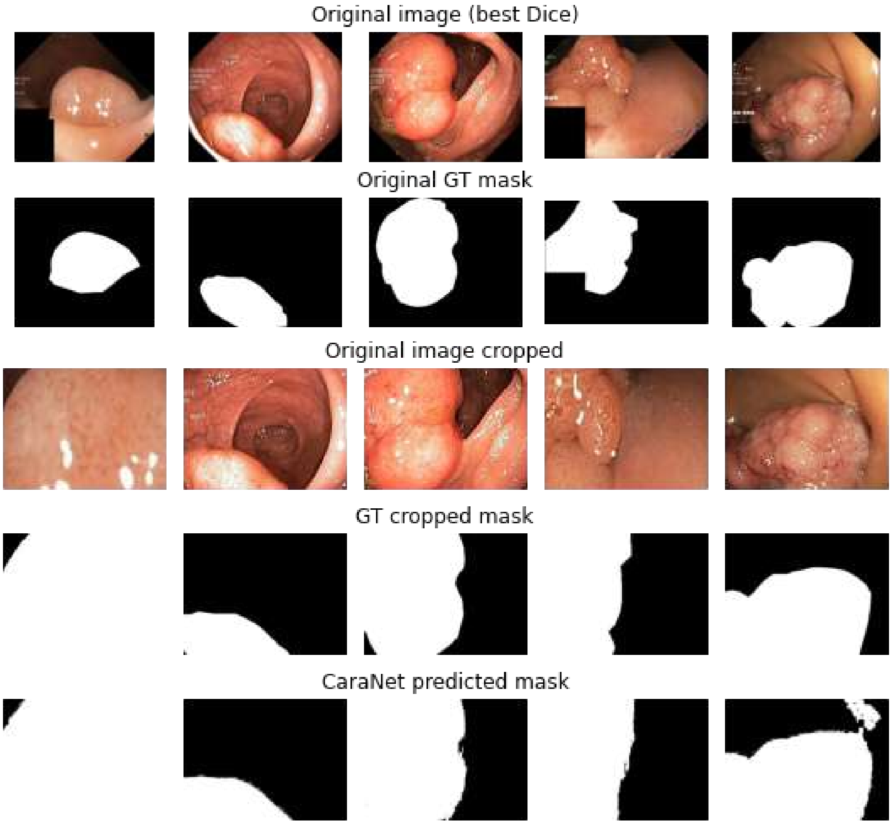

FIGURE 5.1: The best segmentation results

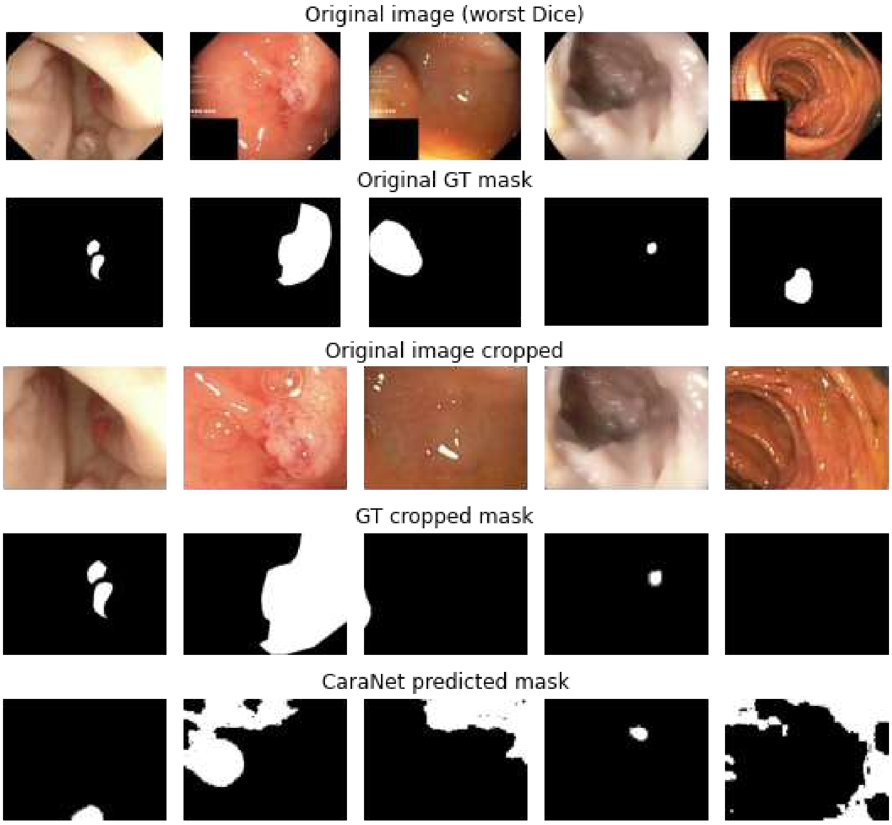

FIGURE 5.2: The worst segmentation results

5.3 Experiments with CaraNet architecture

In this section, all experiments were conducted on HyperKvasir segmented images dataset only.

Partial Decoder CaraNet uses a parallel partial decoder to generate the high-level semantic global map and a set of context and axial reverse attention operations to detect global and local feature information. Firstly, it uses Res2Net as a backbone network to extract low- and high-level features. Then it applies a parallel partial decoder to aggregate high-level features. The partial decoder feature is computed by $PD = p\_d(x4\_rfb, x3\_rfb, x2\_rfb)$. The hypothesis was to test if adding $x1\_rfb$ as an additional input to partial decoder improves the performance. Result Dice score increased a little, precision and FPR metrics improved too, but recall slightly dropped.

In this experiment, only original images with polyps (non empty mask) from HyperKvasir segmented images dataset were used.

TABLE 5.5: CaraNet experiment with Partial Decoder

№	Experiment	Dataset	Dice	Precision	Recall	FPR
9	Original CaraNet	HyperKvasir	0.8782	0.8769	0.9166	0.0237
10	**CaraNet PD input**	**HyperKvasir**	**0.8790**	**0.8918**	**0.8993**	**0.0188**

Gated axial attention Next experiments were inspired by Medical Transformer (Valanarasu et al., 2021) approach.

Authors of Medical Transformer claim that for small-scale datasets tasks, which is often the case in medical image segmentation, the positional bias is difficult to learn and hence will not always be accurate in encoding long-range interactions. In the case where the learned relative positional encodings are not accurate enough, adding them to the respective key, query and value tensor would result in reduced performance. So they proposed a modified axial-attention block that can control the influence positional bias can exert in the encoding of non-local context. With the proposed modification the self attention mechanism applied on the width axis creates gating mechanism which control influence of the learned relative positional encodings have on encoding non-local context. Typically, if a relative positional encoding is learned accurately, the gating mechanism will assign it high weight compared to the ones which are not learned accurately. (Valanarasu et al., 2021)

As in our task we also usually work with typically small-scale images, our model's performance might be also not that good. So we replaced CaraNet axial attention layers with Gated Axial Attention layers taken from Medical Transformer approach. This modification required fixed image size during the training, so we fixed in code $size\_rates = [1]$ and tested $image\_size$ values 320 and 384, which are closest to initial 350 (experiments 11-12).

We also tried architecture with several sequential blocks (experiment 13) and added extra Gated Attention layer to global features from Res2Net (experiment 14). The last experiment gave the best result among all experiments with adding gated axial attention to CaraNet architecture, and even very slightly improved results compared to original, but this improvement can't be called notable (Table 5.6).

The possible reason of that can be lack of LoGo (Local branch+ Global branch) training, proposed in Medical Transformer approach. We haven't added it to out

TABLE 5.6: CaraNet experiments with Gated Axial Attention

№	Experiment	Dataset	Dice	Precision	Recall	FPR
11	CaraNet with Gated Attention (320)	HyperKvasir	0.8707	0.8898	0.8873	0.0178
12	CaraNet with Gated Attention (384)	HyperKvasir	0.8585	0.8600	0.8979	0.0258
13	CaraNet with Gated Attention Sequential Blocks (320)	HyperKvasir	0.8663	0.8596	0.9081	0.0272
14	CaraNet with Gated Attention Sequential Blocks (320) on global features	HyperKvasir	0.8784	0.8847	0.9073	0.0228

CaraNet training. CaraNet doesn't use patches during training, and parallel branch with patches and gated axial attention could make an effect on model's performance.

TABLE 5.7: CaraNet architecture experiments on all datasets

№	Experiment	Dataset	Dice	Precision	Recall	FPR
15	CaraNet PD input	**All**	**0.9086**	**0.9238**	**0.9150**	**0.0080**
16	CaraNet with Gated Attention Sequential Blocks (320) on global features	All	0.8973	0.8981	0.9179	0.0142

We also run two best experiments - with partial decoder input and the last one with gated axial attention - on all datasets. Results prove stability of this approach, but no significant improvement compared to original CaraNet architecture was achieved. Metrics can be seen in Table 5.7.

The best and the worst (using Dice metric) segmentation results of experiments 15 and 16 can be seen in Figures 5.3, 5.4, 5.5, 5.6.

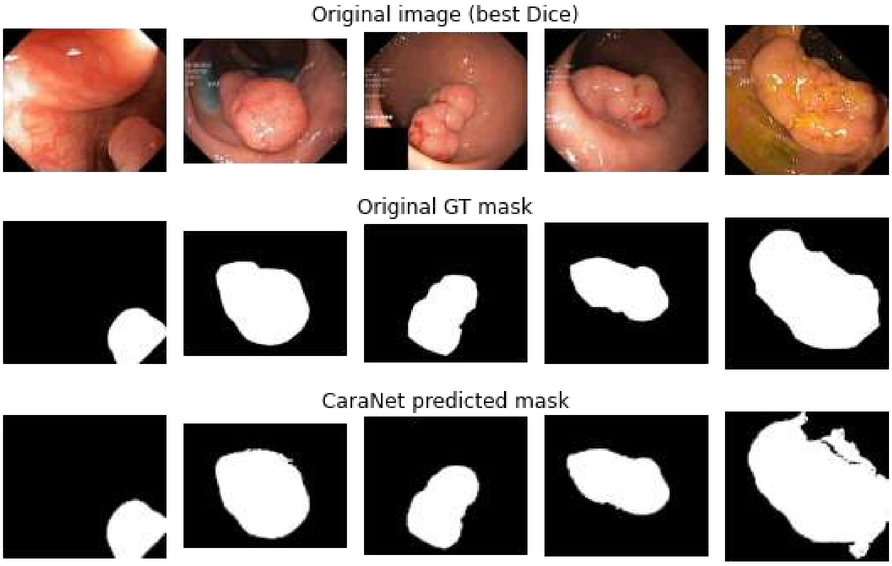

FIGURE 5.3: The best segmentation results (PD input)

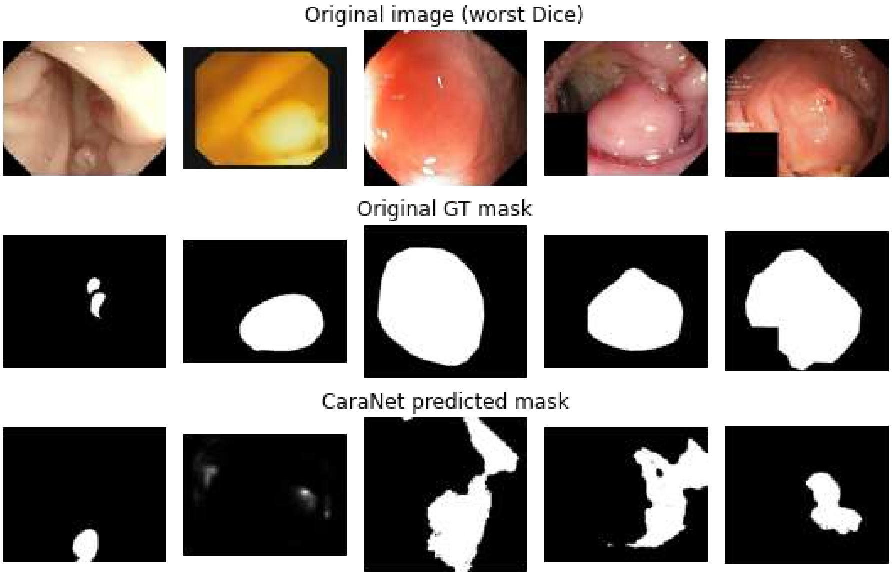

FIGURE 5.4: The worst segmentation results (PD input)

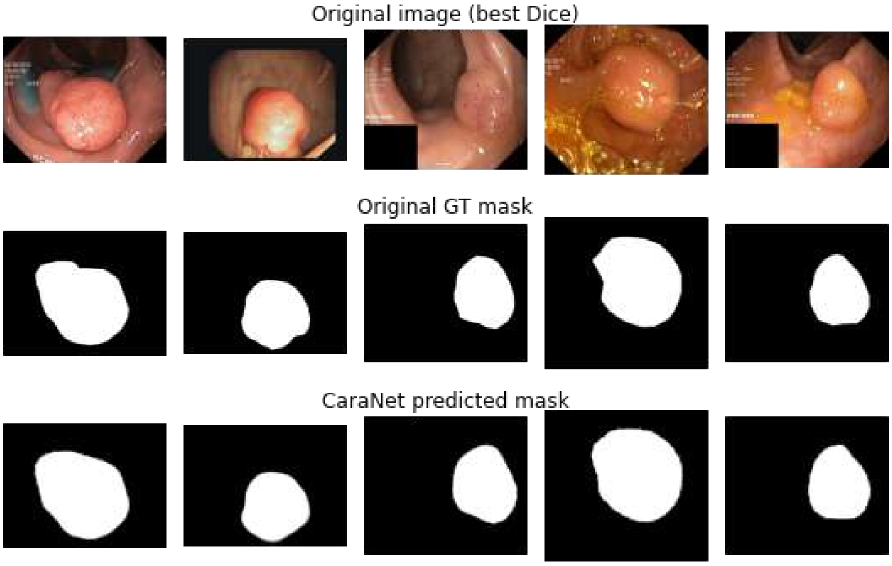

FIGURE 5.5: The best segmentation results (gated axial attention)

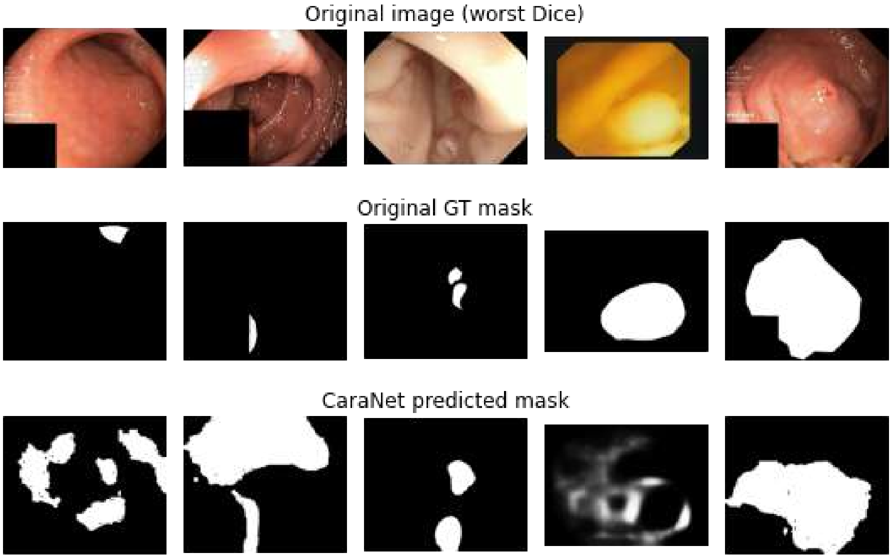

FIGURE 5.6: The worst segmentation results (gated axial attention)

5.4 EndoCV2022 challenge experiments

For this challenge we decided to choose MaskFormer (Cheng et al., 2021b) as the primary model for our approach.

Since we used MaskFormer for binary segmentation, most ground truth classes for each query will be zero, and the cross-entropy loss will rapidly converge to zero. Therefore we changed cross-entropy loss to focal loss to mitigate class imbalance in the classification.

Whereas the challenge dataset is rather small compared to large datasets as ADE20k and COCO-Stuff-10k model was designed for originally, some model hyperparameters were changed to increase the performance and generalizability of our model. Among other, we decreased the number of queries from 100 to 50, FC layers dimensionality from 2048 to 24, and pixel- from 256 to 64. We used a standard convolution ResNet backbone (R50 with 50 layers) instead of SWIN. Normalization coefficients were also recalculated for the PolypGen dataset.

We compared this approach against default CaraNet model settings (Lou et al., 2022, Lou and Loew, 2021). On this challenge training data (validation set), proposed MaskFormer solution has shown better performance. Despite this, in the second test round (test II) MaskFormer approach has shown significantly worse performance. According to leaderboard and workshop rules, participants are not allowed to visualise test samples, so during the challenge we were not able to visually analyze why results on data from this test set worsened so much. Metrics can be found in Table 5.8.

Examples of generated segmentation masks are shown on Figure 5.7.

We also experimented with boundary loss, which showed promising results in other medical imaging tasks, but didn't work well in our case. Detailed discussion on this and other experiments conducted, and impact of TTA (test time augmentation) and CCA (connected-component analysis) during postprocessing can be found in our paper for this challenge (Kokshaikyna, Yelisieiev, and Dobko, 2022).

We assume that including sequence information as an input to MaskFormer (Cheng et al., 2021b) can potentially improve the results. One of the options to do this is to use a Mask2Former (Cheng et al., 2021a) model, which was created for video segmentation and inspired by MaskFormer. Mask2Former is based on Masked-attention Mask Transformer for universal image and video segmentation. It is possible to incorporate their idea in combining the images from the same sequence into a single input with additional dimension responsible for time frames.

Method	Data	Dice	Dice std	FNR
CaraNet	val	0.37516	0.31954	0.71444
MaskFormer	val	0.73587	0.30823	0.28758
MaskFormer	test II	0.5497	0.4319	0.556

TABLE 5.8: MaskFormer model's results on PolypGen2.0 subchallenge

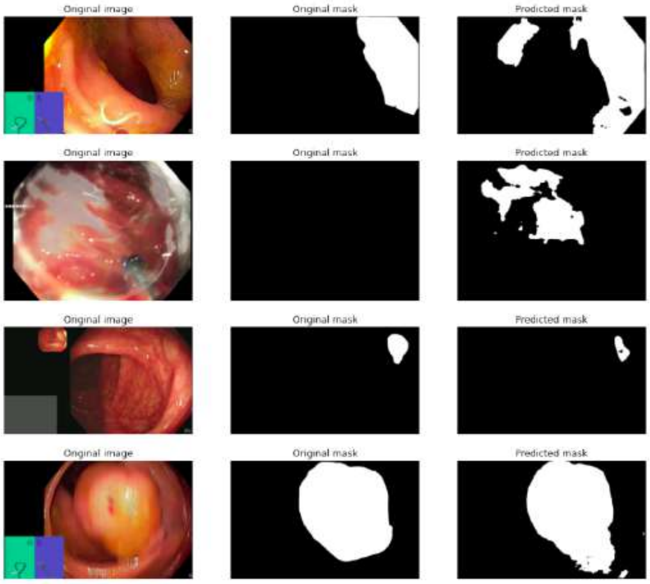

FIGURE 5.7: Examples of MaskFormer model segmentation masks on PolypGen2.0 dataset

Chapter 6

Conclusions and discussion

Conclusions on results

In this master thesis, we made an overview of the existing state-of-the-art approaches on polyps detection and segmentation, selected one of them as a baseline, and explored its robustness on public and custom data.

Unfortunately, we were not able to achieve adequate results on custom data. The most probable reason for that is the difference between public datasets and real-world data. We can state that current model is not robust to all data - to varying degrees, it underperforms for ETIS-LaribPolypDB, second part of the test data from EndoCV 2020 challenge, and our custom dataset.

We also tried a few modifications of the model architecture. These modifications helped to improve the model's performance slightly but did not bring tremendous improvement.

Achieved results with metrics and visualizations provided in the work.

Further work

Further work on this research may include work closer with custom data, ablation study on the influence of augmentations, investigating difference between public data and real world data, collecting more images and annotating.

We also recently got access to PICCOLO dataset[1], which we haven't tried to use in training and testing process yet.

Architecture modifications may include experiments with encoder features (i.e. trying different encoders), combing CaraNet and TransFuse approaches.

One more important thing, which we didn't mention previously, is model explainability, which is important for healthcare domain.

And after achieving desired performance, we would like to target real-time segmentation from endoscopy video frames.

[1] https://www.biobancovasco.org/en/Sample-and-data-e-catalog/Databases/PD178-PICCOLO-EN.html

Bibliography

Akcay, Samet, Amir Atapour Abarghouei, and Toby P. Breckon (2018). "GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training". In: *CoRR* abs/1805.06725. arXiv: 1805.06725. URL: http://arxiv.org/abs/1805.06725.

Akçay, Samet and Amir Atapour-Abarghouei andToby P. Breckon (2019). "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Ali, Sharib (2019). *Endoscopy Artefact Detection (EAD) Dataset*. DOI: 10.17632/C7FJBXCGJ9.1. URL: https://data.mendeley.com/datasets/c7fjbxcgj9/1.

Ali, Sharib et al. (2020). "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy". In: *Scientific Reports* 10.1, p. 2748. ISSN: 2045-2322. DOI: 10.1038/s41598-020-59413-5. URL: https://doi.org/10.1038/s41598-020-59413-5.

Ali, Sharib et al. (2021a). "A deep learning framework for quality assessment and restoration in video endoscopy". In: *Medical Image Analysis* 68, p. 101900. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101900. URL: https://www.sciencedirect.com/science/article/pii/S1361841520302644.

Ali, Sharib et al. (2021b). "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy". In: *Medical Image Analysis* 70, p. 102002. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2021.102002. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000487.

Ali, Sharib et al. (2021c). *PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment*. DOI: 10.48550/ARXIV.2106.04463. URL: https://arxiv.org/abs/2106.04463.

Bernal, Jorge et al. (2015). "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 43, 99—111. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2015.02.007. URL: https://doi.org/10.1016/j.compmedimag.2015.02.007.

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

Cheng, Bowen et al. (2021a). "Masked-attention Mask Transformer for Universal Image Segmentation". In: *arXiv*.

— (2021b). *Per-Pixel Classification is Not All You Need for Semantic Segmentation*. DOI: 10.48550/ARXIV.2107.06278. URL: https://arxiv.org/abs/2107.06278.

Fernández-Esparrach, Glòria et al. (Sept. 2016). "Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps". en. In: *Endoscopy* 48.9, pp. 837–842.

Figueiredo, Isabel et al. (Aug. 2019). "Unsupervised segmentation of colonic polyps in narrow-band imaging data based on manifold representation of images and

Wasserstein distance". In: *Biomedical Signal Processing and Control* 53, p. 101577. DOI: 10.1016/j.bspc.2019.101577.

Gansbeke, Wouter Van et al. (2021). "Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals". In: *CoRR* abs/2102.06191. arXiv: 2102.06191. URL: https://arxiv.org/abs/2102.06191.

Ghosh, Payel et al. (2011). "Unsupervised Grow-Cut: Cellular Automata-Based Medical Image Segmentation". In: *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, pp. 40–47. DOI: 10.1109/HISB.2011.44.

Heresbach, D et al. (Apr. 2008). "Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies". en. In: *Endoscopy* 40.4, pp. 284–290.

Hwang, Jyh-Jing et al. (Oct. 2019). *SegSort: Segmentation by Discriminative Sorting of Segments*.

Jha, D. et al. (2019). "ResUNet++: An Advanced Architecture for Medical Image Segmentation". In: *2019 IEEE International Symposium on Multimedia (ISM)*, pp. 225–255.

Jha, Debesh et al. (2020). "Kvasir-seg: A segmented polyp dataset". In: *International Conference on Multimedia Modeling*. Springer, pp. 451–462.

Jha, Debesh et al. (2021a). "A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation". In: *IEEE journal of biomedical and health informatics*.

Jha, Debesh et al. (2021b). *NanoNet: Real-Time Polyp Segmentation in Video Capsule Endoscopy and Colonoscopy*. DOI: 10.48550/ARXIV.2104.11138. URL: https://arxiv.org/abs/2104.11138.

Kim, Wonjik, Asako Kanezaki, and Tanaka Masayuki (Jan. 2020). "Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering". In: *IEEE Transactions on Image Processing* 29, pp. 1–1. DOI: 10.1109/TIP.2020.3011269.

Kokshaikyna, Mariia, Yurii Yelisieiev, and Mariia Dobko (2022). *Mask Classification-Based Method For Polyps Segmentation and Detection*. URL: http://ceur-ws.org/Vol-3148/paper6.pdf.

Liu, Ze et al. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. DOI: 10.48550/ARXIV.2103.14030. URL: https://arxiv.org/abs/2103.14030.

Lou, Ange and Murray Loew (2021). "CFPNET: Channel-Wise Feature Pyramid For Real-Time Semantic Segmentation". In: *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1894–1898. DOI: 10.1109/ICIP42928.2021.9506485.

Lou, Ange et al. (2022). "CaraNet: context axial reverse attention network for segmentation of small medical objects". In: *Medical Imaging 2022: Image Processing*. Vol. 12032. International Society for Optics and Photonics. SPIE, pp. 81 –92. DOI: 10.1117/12.2611802.

Rother, Carsten, Vladimir Kolmogorov, and Andrew Blake (Aug. 2004). "GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts". In: *ACM Trans. Graph.* 23, pp. 309–314. DOI: 10.1145/1186562.1015720.

Salman, Nassir, Bnar Ghafour, and Gullanar Hadi (Apr. 2015). "Medical Image Segmentation Based on Edge Detection Techniques". In: *Advances in Image and Video Processing* 3. DOI: 10.14738/aivp.32.1006.

Silva, Juan S. et al. (2014). “Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer”. In: *International Journal of Computer Assisted Radiology and Surgery* 9.2, pp. 283–293. DOI: 10.1007/s11548-013-0926-3. URL: https://hal.archives-ouvertes.fr/hal-00843459.

Srivastava, Abhishek et al. (2021). “MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation”. In: DOI: 10.48550/ARXIV.2105.07451. URL: https://arxiv.org/abs/2105.07451.

Sun, Xinzi et al. (2021). *Colorectal Polyp Detection in Real-world Scenario: Design and Experiment Study*. DOI: 10.48550/ARXIV.2101.04034. URL: https://arxiv.org/abs/2101.04034.

Valanarasu, Jeya Maria Jose et al. (2021). *Medical Transformer: Gated Axial-Attention for Medical Image Segmentation*. DOI: 10.48550/ARXIV.2102.10662. URL: https://arxiv.org/abs/2102.10662.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

Vázquez, David et al. (July 2017). “A benchmark for endoluminal scene segmentation of colonoscopy images”. en. In: *J. Healthc. Eng.* 2017, p. 4037190.

Zhang, Yundong, Huiye Liu, and Qiang Hu (2021). *TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation*. DOI: 10.48550/ARXIV.2102.08005. URL: https://arxiv.org/abs/2102.08005.

Zhou, Bolei et al. (2017). “Scene Parsing through ADE20K Dataset”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou, Bolei et al. (2019). “Semantic understanding of scenes through the ade20k dataset”. In: *International Journal of Computer Vision* 127.3, pp. 302–321.