Self-supervised learning methods for medical image analysis in clinical histopathology

A thesis submitted in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE (RESEARCH)

in

Electrical Engineering

by

Ajey Pai Karkala

Entry No. 2019EEY7534

Under the guidance of

Dr. Tapan Gandhi

Department of Electrical Engineering,

Indian Institute of Technology, Delhi

July 2021

CERTIFICATE

This is to certify that the work contained in this thesis titled **"Self-supervised learning methods for medical image analysis in clinical histopathology"** by **Ajey Pai Karkala** (2019EEY197534) for the award of **Master of Science (Research)** in **Electrical Engineering** from the **Department of Electrical Engineering, Indian Institute of Technology, Delhi** is a bona fide work carried out by him under our guidance and supervision. The matter submitted in this dissertation has not been submitted for an award of any other degree or diploma anywhere unless explicitly referenced.

Signed:

Dr. Tapan Gandhi

Associate Professor

Department of Electrical Engineering

ACKNOWLEDGEMENT

Firstly, I'd like to thank Dr. Prathosh AP and Dr. Mausam for taking me on as a student to work on this project. It is the single most important event which has altered the course of my entire life. Even though I did not have a computer science background, Dr. Prathosh didn't hesitate to give me a chance to work and develop my skills. The theoretical intuition that he imparted through lectures or during casual discussions is something I will take forward. The weekly meetings with Dr. Mausam helped me immensely in developing a scientific acumen that made most of this work possible. His inputs regarding the project and tips on how to sustain in campus have helped me beyond measure.

I'd specially like to thank Dr.Prasenjit Das and Dr.Govind Makharia from AIIMS, New Delhi. Owing to the inter-disciplinary nature of this project, Dr.Prasenjit spent enormous hours with me poring over the details of intestinal anatomy and the salient features of Celiac Disease for which, I'm extremely thankful. I thank the members of his Pathology lab – Mr.Mukesh and Ms.Shreya for making me feel welcome at AIIMS. I'll always cherish the time I spent in that lab forever. Ms.Lalitha is specially mentioned for her efforts in helping me annotate the histological images due to which, the novel dataset became a reality.

I will forever be grateful to Dr. Tapan Gandhi for his support during the final months of my thesis. No words can do justice for the mentorship he's rendered me.

I thank Nisarg Bhatt for his contributions to this work and his collaborative role with me during this project.

I'd like to thank my spiritual masters Smt. Jayanti Prabhu and Sri. Vishwanath Shastry for anchoring me in my failure and in doubt.

Most importantly, I'd like to thank my parents and my brother for believing in me and providing me with the space, time and resources to pursue this degree at IIT-Delhi.

I dedicate this body of work solely to Lord Dattatreya and profusely prostrate at his lotus feet.

Ajey Pai Karkala

Date: 27 August 2022

ABSTRACT

Self-supervised learning is emerging as a promising class of unsupervised learning methods that make the best use of available data for representation learning. Specifically in the medical imaging domain - due to the limited availability of large amounts of annotated data, self-supervised learning can help pave way to successful applications of deep learning methods for clinical use. However, most existing self-supervised algorithms employ generic, standalone pretext tasks that aren't specifically designed to learn morphological representations or representations of positional relationships among different tissues present in an image sample. Hence, they learn medically inexplicable representations which renders even state-of-the-art models as black-box systems unreliable for practical clinical use. For reliable applications, clinical interpretability of deep learning systems becomes imperative in medical diagnostics. In this work, along with fully supervised models that perform semantic segmentation and localization of areas of interest, we also propose novel self-supervised learning methods by using clinical knowledge as motivation while designing different pretext tasks for learning reliable representations. We evaluate these methods on a novel dataset of histological images of the human duodenum that our team curated along with expert gastroenterologists, pathologists at AIIMS-New Delhi. We show promising results with these new self-supervised learning approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Self Supervised Learning

Deep convolutional neural networks (CNNs) perform based on the power of CNNs and the amount of training data available. Different models have been developed to increase the power of CNNs, and larger and larger datasets are being collected for a variety of tasks these days. Networks like DenseNet [1], ResNet [2], GoogLeNet [3] ,VGG [4], AlexNet [5] and large scale datasets like ImageNet [6], Coco [7], OpenImage [8] have been proposed for training very deep CNNs.

Since CNNs require massive sets of annotation rich training data to learn discriminative representations, collecting such large sets of well-annotated data is prohibitively expensive due to its tenuous nature specially in the field of medical imaging across different modalities. Fortunately, in recent years, advanced methods on unsupervised learning have been developed that learn very strong representations without the need of human annotated data. One of the subsets of such unsupervised learning methods is called as self-supervised learning (SSL).

Self-supervised learning exploits unlabelled data to learn image abstractions. Specifically in computer vision, the goal of self-supervision is to construct image representations that are semantically meaningful without their semantic annotations. Self-supervision is a good method to initialize model parameters for training annotation-efficient downstream models for image classification, object detection, semantic segmentation and panoptic segmentation.

Pre-training tasks are tasks that have been pre-designed for deep networks to solve. Learning the objective functions for these tasks allows you to learn the visual features (eg: Image reconstruction). These tasks can be predictive tasks, generative tasks, contrastive tasks or different combinations of these individual task types. They must be designed by taking into consideration the properties of images like colour, structure, content and semantics to ensure optimal learning.

Downstream tasks are computer vision tasks that can be used to evaluate the quality of visual features learned by the CNNs by self-supervised learning. Generally, the downstream tasks require human-annotated data for training. To solve these tasks, pre-trained models help immensely when the training data is scarce.

For self-supervision on image data, different pre-training tasks share two common properties:

- By solving these tasks, the deep models must be able to capture visual features in the data.
- The supervisory signals needed to train the models must be generated from the data itself (without human annotations) by leveraging the colours and/or structural properties of images.

Figure 1.1 illustrates the general pipeline for self-supervised learning. Training SSL models typically consists of two stages. The first stage is the pre-training in which the CNNs are tasked with solving a pre-defined task. These tasks are designed such that the CNNs learn visual features from the images by solving them. In that, the initial blocks of the CNN learn kernels that capture general features such as textures, edges, and corners while the deeper blocks learn more fine-grained features that maybe useful for downstream tasks. After the pre-training, the learnt features can be transferred to downstream tasks. Using the self-supervised model as pre-trained model for the downstream tasks is advantageous when the labelled dataset available for the downstream tasks is limited. It improves performance and reduces overfitting.

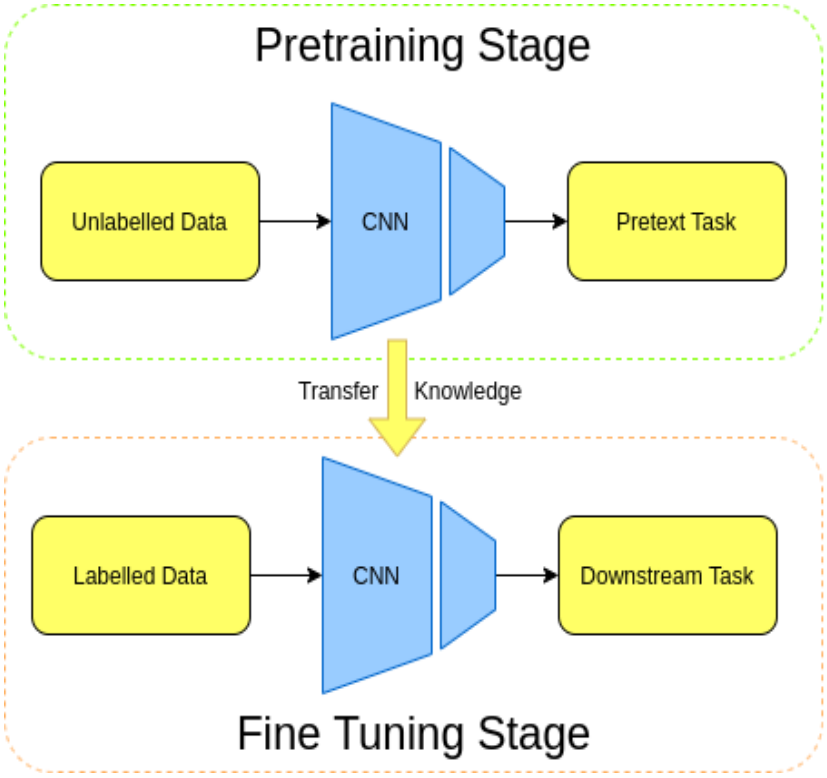


Figure 1.1: A skeletal workflow of self-supervised learning. By training CNNs to complete a pretext task, the visual feature is learned. The learned parameters are then used as a pre-trained model after self-supervised pretext task training is complete. In the fine-tuning phase, these can be applied to additional computer vision tasks.

1.2 Object detection

To understand an image completely, it isn't enough to just classify them. It is important to find objects and locations of each objects in the image. This task of determining the contents and locations of different objects in a given image is called as object detection. It is a coarse prediction task that involves taking an image input and classifying certain regions as belonging to different categories based on visual features. Additionally, regressing bounding box coordinates on the classified objects.

Figure 1.2: Pictorial representation of object detection

1.3 Semantic Segmentation

In modern medical imaging, semantic segmentation is used on 2D images, volumetric images, and videos. It is one of the major issues facing computer vision, and its resolution would improve the ability to better scene understanding. In recent years, increased number of applications of computer vision benefit from scene understanding. This emphasizes how crucial semantic segmentation is as a fundamental idea in computer vision.

Earlier, researchers utilized classical computer vision and machine learning techniques to address this issue. But, with the advancement of deep learning, CNNs are being used to tackle semantic segmentation. These deep and efficient networks are shown to perform faster than traditional methods surpassing them in accuracy by large margins.

The high level task of semantic segmentation includes classification of image pixels into different semantic categories for given image data with different semantic pixel contents.

Some applications of semantic segmentation in computer vision include autonomous driving [9], [10], [11], Human-machine interaction [12], computational photography [13], augmented reality and many more. Figure 1.3 illustrates the task of semantic segmentation.

Generally, CNNs are used for semantic segmentation of biomedical images. These deep networks assign values to each pixel of an image that represents its probability of belonging to a certain semantic class. The probability of a pixel belonging to a certain class may not always be independent of the class to which a different pixel in the image belongs to. They are produced by a set of weights and activations in the network that are learnt by minimizing a loss (Typically Dice loss, Cross-Entropy loss etc). These measure the overlap between

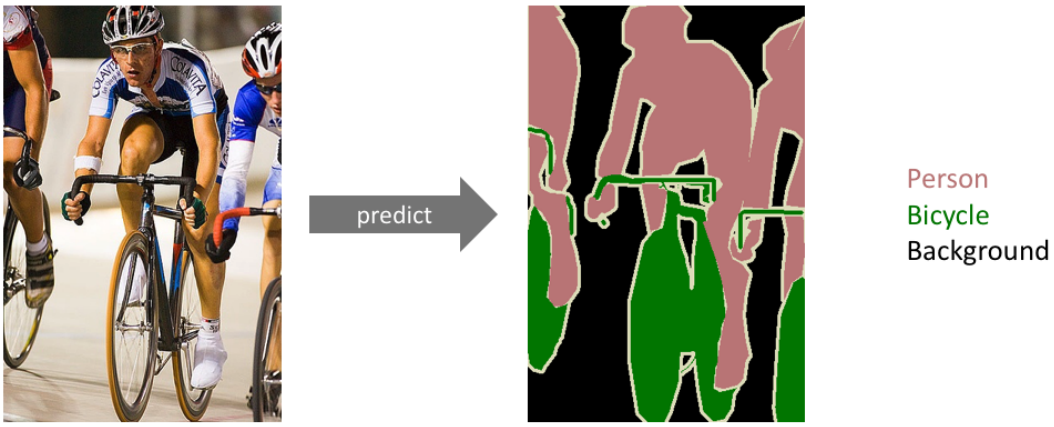


Figure 1.3: Pictorial representation of semantic segmentation.

the ground-truth pixels with the corresponding pixels in the predicted segmentations. While calculating such a loss, each pixel is considered independently.

Since pixel semantics are highly localised and dense in nature, it is particularly a challenging task to introduce global semantics in the training process of CNNs that will produce globally coherent segmentations specifically for medical data with limited annotations and high variability. The different instances of objects belonging to the same class exhibit a high degree of variation in terms of shapes and sizes. For some modalities of medical images and diagnostics, there is often a large degree of inter-observable variability in classifying different objects in a given image even among highly trained and experienced doctors.

1.4 Celaic Disease

Celiac Disease is a chronic systemic autoimmune disorder induced by a protein called gluten in several food substances like wheat and barley. If patients having a predisposition to this disease consume gluten, an immune reaction is triggered in their small intestines. Overtime, this reaction damages the finger-like structures called villi on the inner lining of the small intestine preventing them from absorbing certain nutrients (malabsorption). The damage to the small intestine can cause bloating, diarrehea, fatigue, weight loss and can lead to serious complications.

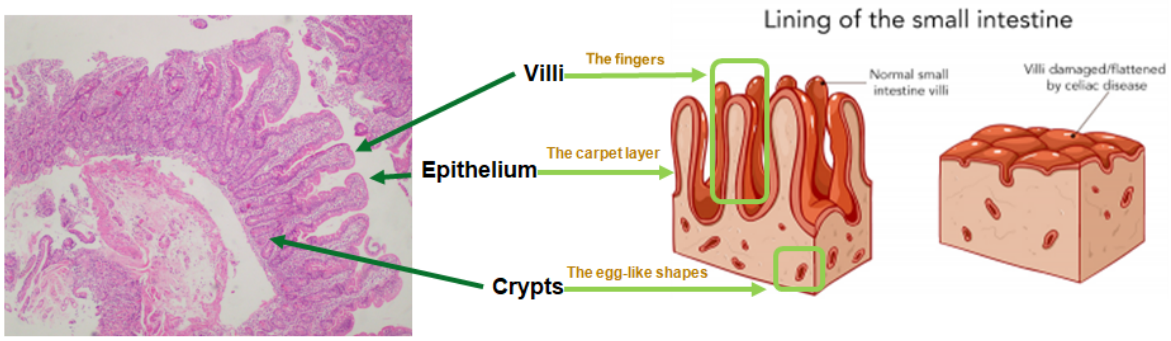


Figure 1.4: Illustration of Celiac Disease. The Villi are finger-like projections as indicated. The epithelial layer is the outer layer of the villi and the egg shaped tissues under the villi are called crypts.

Screening studies in different populations have shown that the prevalence of celiac disease is around 1% or more in both the United States and Europe [14], [15], [16], [17]. This means that, for every patient with the diagnosis of celiac disease, 3–10 remain undetected. Moreover, the prevalence of detected cases of celiac disease is much lower, from 0.27% to 0.02%. In the wheat eating north Indian region, 1 person in every 96 people have celiac disease that largely go undetected [18].

The investigation for Celiac Disease is done by obtaining a biopsy from the small intestine of a patient. Hematoxylin and eosin (H&E) stain is used to prepare histological slides for interpreting the biopsy under a microscope. An expert pathologist looks at these slides and reports their findings based on a particular histological classification protocol and suggests treatments if the patient has the disease. Figure 1.4 illustrates a typical histological image of the human duodenum on the left. Individual tissues in the image are indicated in an animated image on the right for clarity.

Existing histological classification methods for the interpretation of small intestinal biopsies are based on qualitative parameters with high intraobserver and interobserver variations. Recently a group of researchers at the All India Institute of Medical Sciences, New Delhi (AIIMS, New Delhi) proposed a novel quantitative histological (Q-histological) classification system specifically for Celiac Disease. It is shown to be better than existing classification systems [19].

The current work was done in collaboration with researchers at AIIMS. Out of our fruitful collaboration, we have developed a completely new duodenal histopathology dataset that is richly annotated. The technical work focuses on developing deep learning algorithms to assist the clinicians follow the Q-histological rules for accurate detection and grading of duodenal biopsies. Additionally, we explore new self-supervised methods to improve the model performance with unannotated data.

This thesis records the following contributions:

- *Creation of a novel duodenal histopathology dataset*– Careful annotations were done by marking out precise boundaries of different important tissues in a given biopsy. Based on the orientation, shape and integrity of different tissues, areas of clinical interest were marked. All the annotations used for this work have been verified for correctness by the pathologists at AIIMS.

- *A software solution to mark areas of clinical interest*– A fully supervised algorithm was designed to detect areas of clinical interest enabling meaningful interpretation of histological images.

- *Novel self supervised learning methods*– Two novel self-supervised learning methods are proposed for meaningful semantic segmentation using unlabelled histological images.

1.5 Tissue Morphology in Clinical Histopathology

In the field of pathological analysis of histological images of human tissue resections, the tissue morphology is of utmost importance to the clinician in the diagnostic process. In that, there are very specific structural arrangements that are indicative of the underlying cause or mechanism of disease. These structural eccentricities become very pronounced in H&E stained tissue slides and allows a well-trained pathologist to arrive at a diagnosis. The diverse morphologic patterns thus observed under a microscope arise as a result of the intricate biological interplay underlying the specimen's presentation [20]. Moreover, in understanding these morphologies, there lies a potential to directly infer the molecular phenotypes [21]

which are indicative of the genotype that causes a particular disease. Although, the importance of tissue morphologies is widely acknowledged by the histopathologists, its interpretation during clinical diagnostics suffers from intra and inter observer variability with a constant struggle to precisely define and categorize meaningful morphologies [22].

The application of deep learning to the field of computational pathology is an exciting intersection but it begets a very important question for computer scientists - is it possible to represent the symbolic notions of structure, shape and spatial arrangement of tissues using compact codes (as computed by gradient descent) in such a way that allows for easy retrieval and interpretation of the original symbols?

This body of work is an attempt at finding some answers to this question. Then, there are other limitations in this field posed by sparsity of data and/or annotations that need no introduction.

Here, we systematically train a self-supervised learning (SSL) model which specifically incorporates in its learning mechanism, a general notion of spatial arrangement of many meaningful parts that become meaningless when their arrangement is perturbed. While this model learns to reconstruct these parts from a large corpus of histopathology image data, it becomes *familiar* with morphologies that may occur in a variety of H&E stained tissue specimen. This may prove useful while retraining the model for tangible downstream tasks like semantic segmentation of specific tissues in a specimen or classifying the disease grades directly from the histopathology image.

1.6 Organization of the Thesis

Chapter 2 includes the details about the annotation process and the recently proposed Q-histological rules for classifying Celiac Disease. In Chapter 3, supervised baselines are discussed along with performance metrics used to evaluate the model. Additionally, two unique supervised approaches are discussed that propose bounding boxes around the areas of clinical interest by using segmentation features. The chapter is concluded with a discussion on why self-supervised learning methods are useful. In Chapter 4, literature survey of different popular self-supervised learning methods are presented. We justify our approach in

the context of encoding image semantics in representations of medical images. In Chapter 5, we propose three novel self-supervised learning methods. Finally, Chapter 6 summarizes the entire work and also provides the directions for future work.

Chapter 2

Data annotation and the Q-histological classification system

2.1 Data Annotation

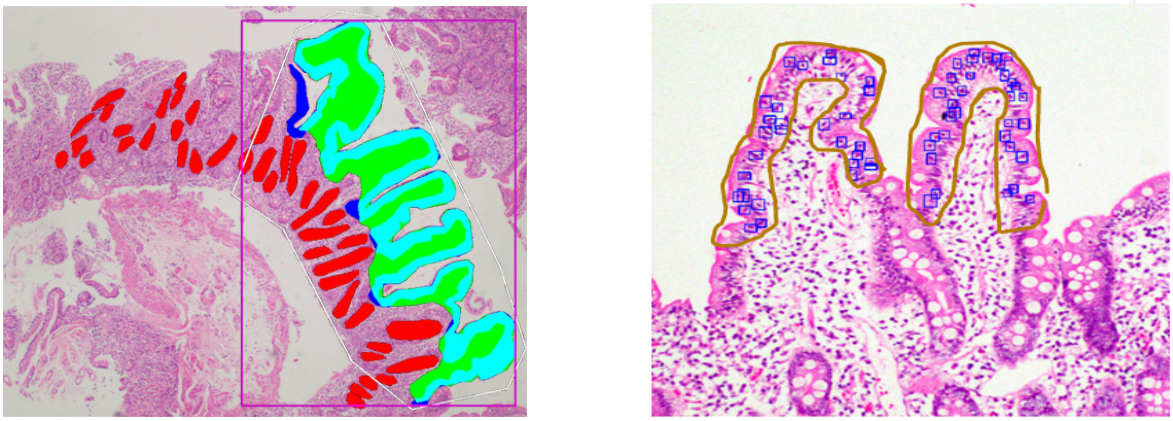
The data collection was approved by the Institutional Ethical Committee (IEC-858 dated 16/12/2019) at the All India Institute for Medical Sciences, New Delhi.

In the first phase of annotation, 1600 histological images were collected from the department of pathology. The images were captured through an Olympus BX50 microscope at 4× zoom using a DP26 camera. Out of these, 800 images have been annotated using the labelme tool. Out of the 800 images, 573 images have been verified for correctness by the pathologists. In each image, Good Villi, Denudated Villi, Good Crypts, Circular Crypts, Epithelium, Brunner's Glands, Muscularis Mucosa and the Interpretable Region were annotated. The histological conditions for an Interpretable Region are explained in the next section. Figure 2.1(a) shows an overlay of annotations on the original histological image. Table 2.1 contains the details of the first phase of annotations.

In the second phase of annotations, images were captured at 20x zoom. The images were taken mostly around the Villi region of the biopsy slide. A total of 65 images were annotated. In each image, Intra-Epithelial Lymphocytes (IEL) and Epithelial Nuclei were annotated. Refer Table 2.2 for details. Figure 2.1(b) shows an example of IEL annotations in the Epithelial region.

S.No	Tissue Name	Total Annotated
1.	Good Crypts	10912
2.	Circular Crypts	10595
3.	Good Villi	714
4.	Denudated Villi	787
5.	Epithelium	1957
6.	Muscularis Mucosa	618
7.	Brunner's Gland	262
8.	Interpretable Region	479
9.	Slanted Villi	28

Table 2.1: Annotation details of 573 verified images. These images were used to train the cascaded network which performs tissue segmentation and localizes regions of clinical importance.



(a) (b)

Figure 2.1: (a) Green - Good Villi. Red - Crypts. Cyan(or blue) - epithelial layer. The bounding box denotes the Area of clinical interest or the Interpretable Region. (b) IELs annotated using blue bounding boxes. Brown borders denote the epithelial regions at the tip of Good Villi.

S.No	Tissue Name	Total Annotated
1.	Intra-Epithelial Lymphocyte	2090
2.	Epithelial Nuclei	6518
3.	Epithelial Area	94

Table 2.2: Second phase annotation details. 65 images were annotated at a deeper resolution to mark Epithelial area, IELs and Epithelial Nuclei.

2.2 The Q-histological classification rules for grading biopsies of small intestine

Existing histological classification systems for assessing Celiac Disease in the small intestine are extremely qualitative in nature. Due to this, there exists a large inter-observer variability among pathologists in most diagnostic parameters except for Intra-epithelial lymphotcyte count. Recently, [19] proposed a quantitative classification system which was shown to be the better than earlier systems. There is very low inter-observer disagreements in this method and thus, ideal for clinical application. Since the nature of the diagnostic method is quantitative, using computational methods to analyse biopsies and measure the diagnostic parameters becomes possible. Table 2.3 lists the quantitative parameters corresponding to different grades of celiac disease.

Class	IEL count /100 Enterocyte cells	Villous Height Change	Cd:Vh ratio
Type 0	< 25	–	–
Type 1	≥ 25	0.7	< 0.5
Type 2	≥ 25	≤ 0.7	≥ 0.5
Type 3	≥ 25	≤ 0.7	≥ 0.5

Table 2.3: The Q-histological parameters for Celiac Disease classification (Duodenum)

2.3 Discussion

Due to the highly quantitative nature of the problem, computer vision algorithms can be used to measure the parameters used for grading biopsies. More concretely, given a biopsy image, our algorithm must accurately:

- Segment out important tissues like Good Villi, Good Crypts and Epithelium.
- Mark areas of interpretability with bounding boxes in regions where all the three mentioned above are co-located.
- Filter out areas where the co-location of these tissues exits but an additional Brunner's Gland is present.
- Measure the heights of each Good Villous and depths of each Good Crypt present in the area of interpretability.
- Count the IELs present in the Epithelial region of the Good Villi.

Chapter 3

Supervised learning: Baselines and methods

3.1 Introduction

The algorithm development was undertaken in stages. The first stage of the project was to develop a model that could perform semantic segmentation of tissues and propose areas of clinical interpretability. Figure 3.1 illustrates this problem pictorially. Given an input image, a segmentation model has to accurately identify individual tissues. Additionally, it should automatically propose areas of clinical interest based on their co-location. Several approaches were designed to perform this task. This chapter records them in detail.

3.2 The Fully Supervised Baseline

We compared U-Net [23], ResU-Net++ [24] and Attention U-Net [25] on the segmentation task of three semantic classes of the Duodenal histology dataset - Good Crypts, Good Villi and Epithelium. Attention U-Net empirically performed better than the other two. Hence, it is chosen as the baseline model for our experiments. It is a fully convolutional network with an encoding path and a decoding path with Attention Gates connected with skip connections at corresponding layers of each paths. The performance comparison with other segmentation models of the Attention U-Net on three classes is given in Table 3.1.

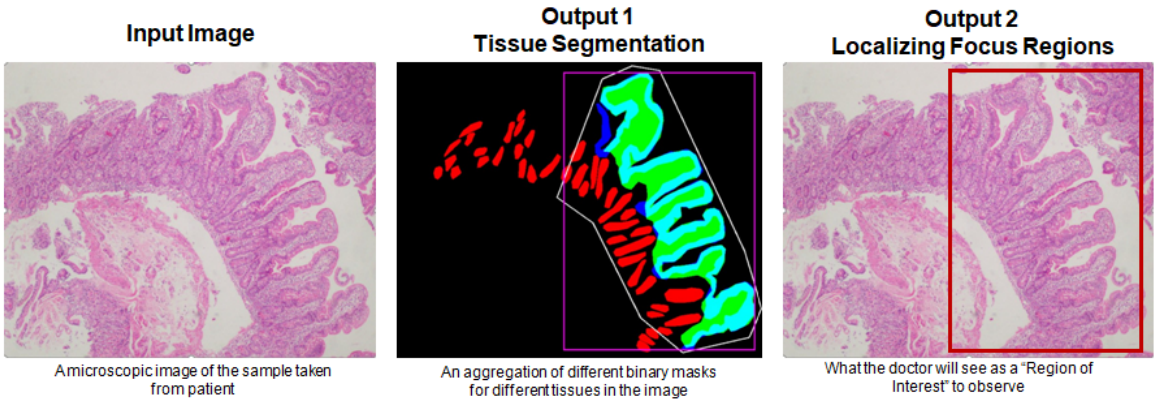


Figure 3.1: Pictorial representation of semantic segmentation and localization of areas of interpretability.

Model	Good Crypts	Good Villi	Epithelium	Average
U-Net [23]	50	51	56	52
ResU-Net++ [24]	53	55	57	55
Attention U-Net [25]	59	57.7	61	59

Table 3.1: Performance comparison of various baseline models on three classes of the Duodenal Histology dataset. Reported scores are Dice Coefficients on 60 test images.

3.3 Dataset, Loss Function and Evaluation Metric

We used 300 fully annotated histological images for training the fully supervised models. A separate batch of 40 images were used for validation during training. For model evaluation, additional 60 images were used. While training the cascaded model, the classes Good Crypts and Circular Crypts were combined into a single class - Crypts. The cascaded model was trained to segment Good Villi, Crypts, Epithelium and Brunner's Gland. Whereas, the segmentation path in the Joint Learner method segmented only Good Villi, Good Crypts and Epithelium and used the intermediate features in the decoder for bounding box regression. For both the models, the performance metric used for model evaluation was the Dice Coefficient given in Eq 3.3. For the cascaded model, among different loss functions tried, we found that the Focal Tversky loss performed the best at appropriately weighting highly unbalanced classes like Brunner's Gland. Empirically, it was also the best loss function to avoid False

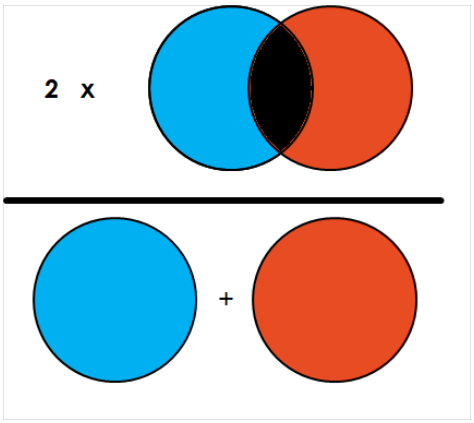


Figure 3.2: visual illustration for calculating Dice Score.

Negatives since it gives a higher weightage to them during the loss calculation. The Focal Tversky loss is given by:

$$FTL = (1 - TI)^{\gamma} \tag{3.1}$$

Where $\gamma = 1.0$ and TI is the Tversky Index as calculated by Equation 3.2 with $\alpha = 0.7$

$$TI = \frac{TP}{TP + \alpha FN + (1 - \alpha)FP} \tag{3.2}$$

The evaluation metric used to test the performance of the models was Dice Coefficient. Figure 3.2 illustrates the metric visually. In statistical terms, it is called the F1 score. given two sets A and B, the Dice coefficient can be written as:

$$\textit{Dice coefficient} = \frac{2|A \cap B|}{|A| + |B|} \tag{3.3}$$

Dice Score for Object Detection: Popularly, the Mean Average Precision (mAP) metric is used to evaluate object detection models. However, since our task is to identify tissue regions inside bounding boxes, we employ the Dice Score to evaluate object detection quality in a special way. Figure 3.3 illustrates this method clearly. Once an object detection model predicts bounding boxes, we post process the regions inside them to only retain pixels that belong to the biopsy by discounting the background. Similar post processing is done for regions inside ground truth boxes. Then, the Dice score is measured between corresponding boxes. This makes the evaluation process precise and accurate.

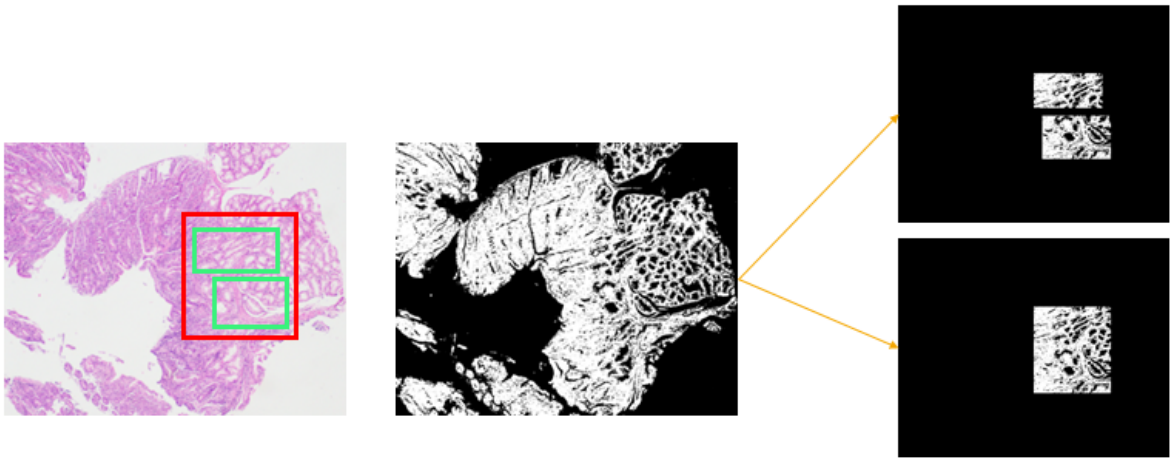


Figure 3.3: visual illustration for evaluating object detection using the Dice Score. The red box is the ground truth. The green boxes represent the output predictions from an object detection algorithm. We process the whole image to retain only those pixels belonging to the biopsy and mask the background. Regions inside individual boxes are cropped and the Dice Score is calculated.

3.4 Methods

We propose two fully supervised methods to perform the task of tissue segmentation and localization of interpretable areas. We begin our design by understanding the relationships between tissues in the biopsy and their relevance for marking the area of interpretation. The area of interpretation should contain pronounced villi structures, elongated crypts that are oriented towards the villi and intact epithelial lining on the periphery of the villi.

3.4.1 The Joint Learner

Our first method is called the Joint Learner. It uses these spatial relationships to propose areas of interpretation by tapping features from the intermediate layers of the segmentation decoder. The resulting model becomes interpretable in the sense that, the bounding boxes proposed by the model has a dependency on the segmentation of important tissues and their implicit relationships. Figure 3.4 shows the block diagram of the Joint Learner.

In contrast to blackbox Deep Learning models that take image inputs to provide bounding box predictions like FasterRCNN [26] , Yolov3 [27] etc, the Joint Learner incorporates some

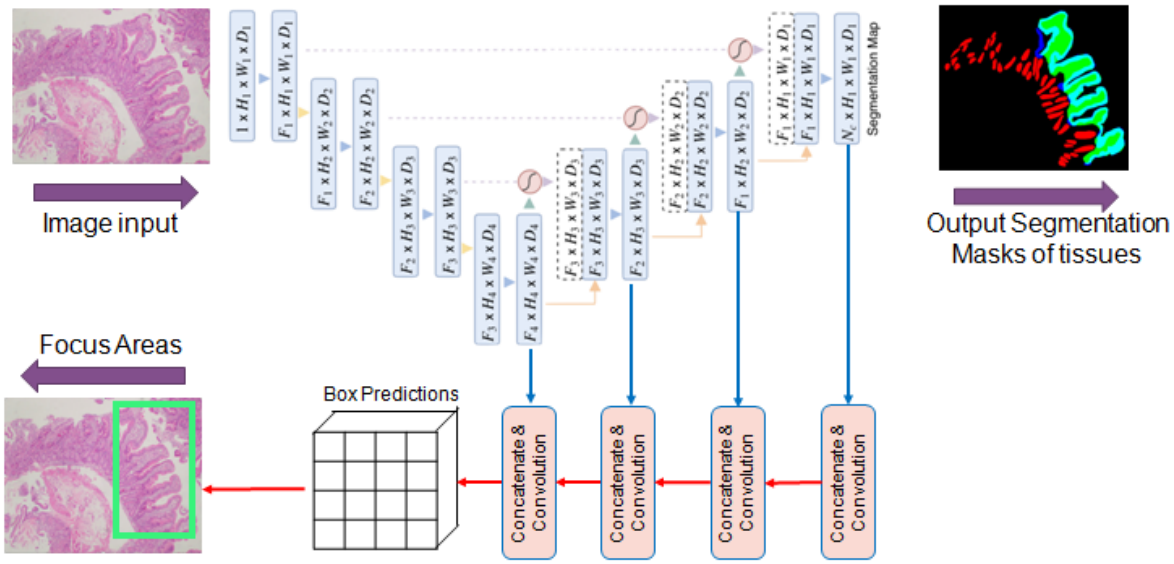


Figure 3.4: Model architecture for the Joint Learner. Features from the decoder of the Attention U-Net are tapped and used to regress bounding boxes.

Model	Dice Coefficient
EfficientDet [28]	0.66
Yolov3 [27]	51
FasterRCNN [26]	54
Joint Learner	0.56

Table 3.2: Performance comparison of Joint Learner with prior arts.

intuition into the learning process that is clinically relevant. Therefore, outputs of this model are relatively more explainable. The performance comparison of the Joint Learner with different localization models is given in Table 3.2.

Although EfficientDet [28] performs better at localising the interpretable regions, the Joint Learner's interpretability and its upward performance in comparison with Yolov3 and FasterRCNN is promising. Incorporation of clinical knowledge into the learning mechanism of deep learning models is of paramount importance. The implementation of this method can be found here. Some output images from the Joint Learner are shown next.

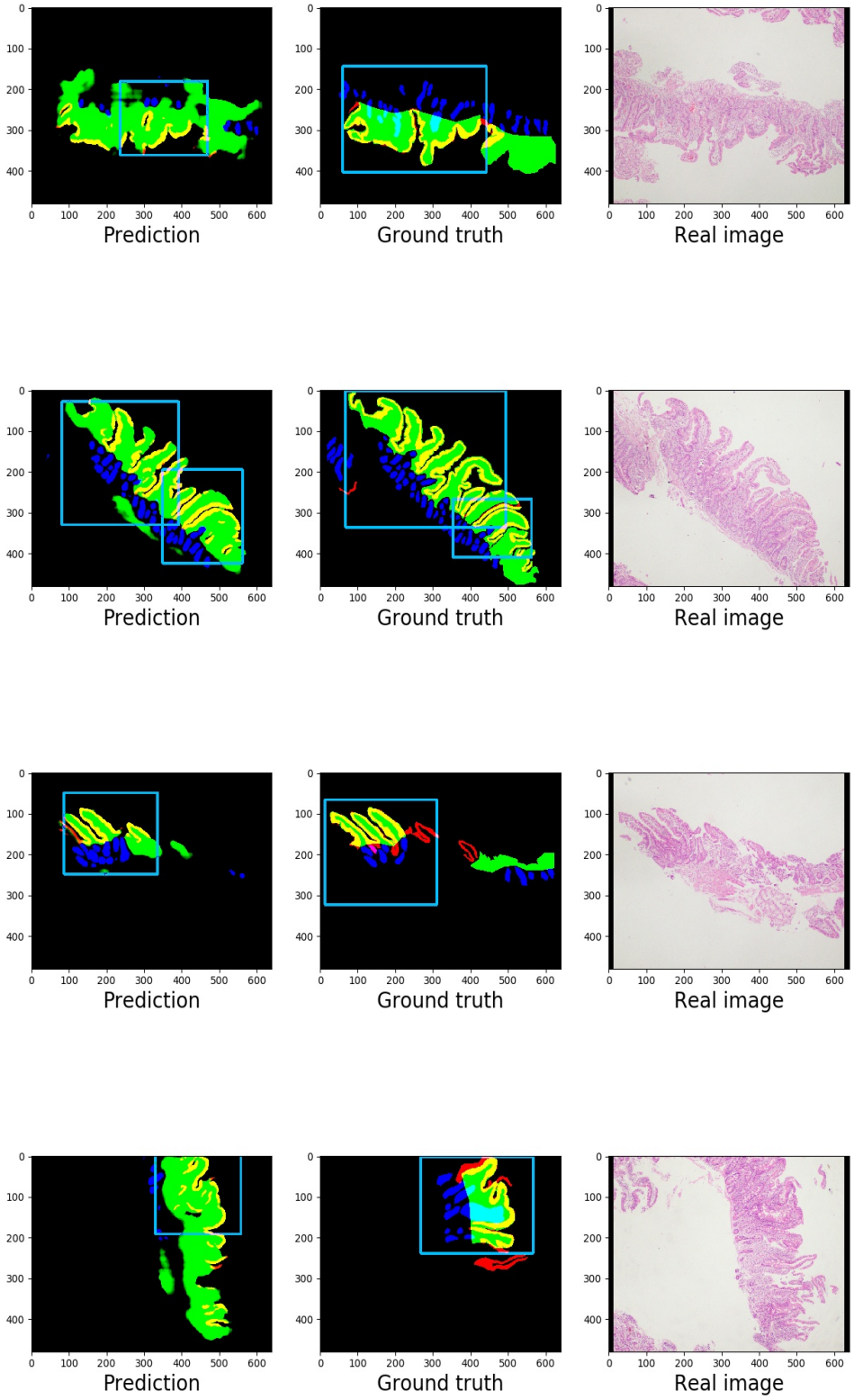


Figure 3.5: Some outputs from the Joint Learner Network

3.4.2 The cascaded model

The second method we propose is a cascaded system of two segmentation models and a localization model as shown in Figure 3.6. Given a histological image - firstly, a segmentation network identifies important tissues like Villi, Crypts and Epithelial layers using an Attention-Unet. Then, another segmentation network segments the edges of crypts. These edges are used in circularity analysis of crypts to differentiate between Good Crypts and Circular Crypts. Secondly, the segmentation outputs after filtering out Circular Crypts are fed to an Efficientdet and localization of regions with clinical importance is done. The region proposals along with tissue segmentations are at par with human-level performance. The performance metrics for image segmentation of the cascaded model on different data-splits is given in Table 3.3. The average Dice score obtained on the test set for localization is **54.52%**. The implementation can be found here.

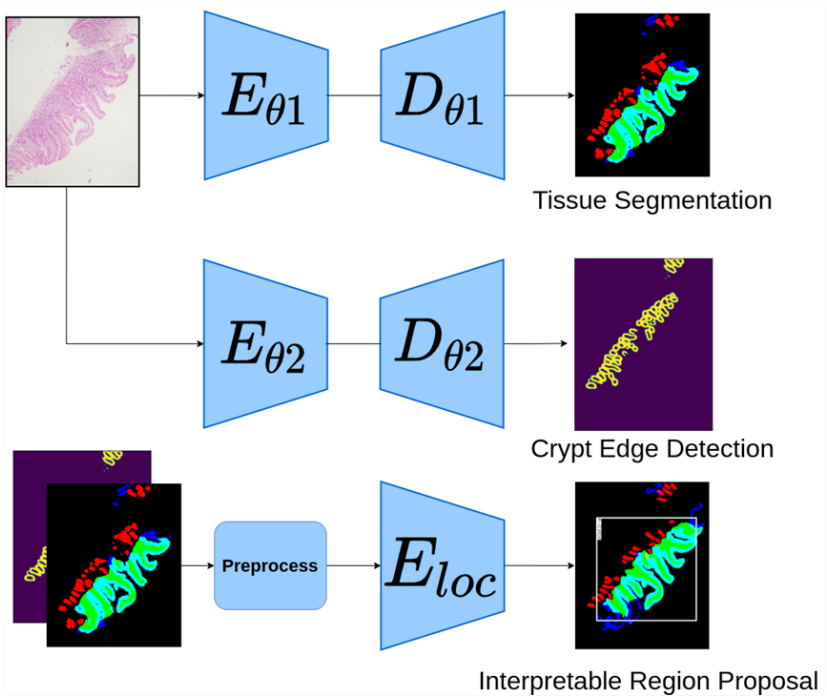


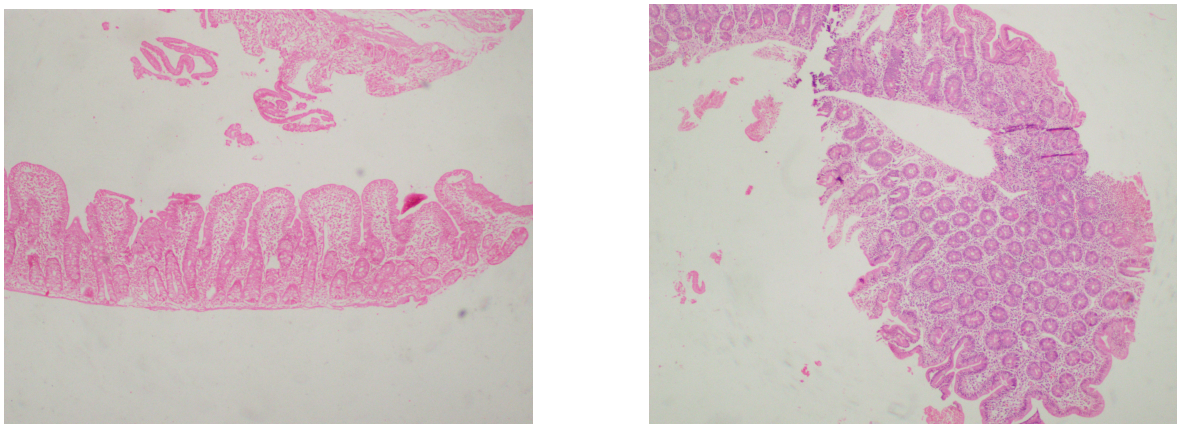
Figure 3.6: Cascaded system for tissue segmentation and bounding box regression. The segmentation outputs from $E_{\theta 1}$ and $D_{\theta 1}$ along with detected crypt edges from $E_{\theta 2}$ and $D_{\theta 2}$ after preprocessing are used for Bounding box regression using an EfficientDet (E_{loc}).

Circularity analysis of Crypts: Crypts are relatively small, densely packed structures in

Split name	Good Crypts	Good Villi	Epithelium	Brunner's Gland	Average Dice
Train	71.1%	71.4%	69.7%	91.1%	75.8%
Validation	67.4%	58%	67.7%	86.3%	69.8%
Test	68%	60.3%	63.8%	85.3%	69.3%

Table 3.3: Dice scores of the cascaded system on different data splits. We used 60 images for testing, 40 images for validation and 300 images for training our algorithm.

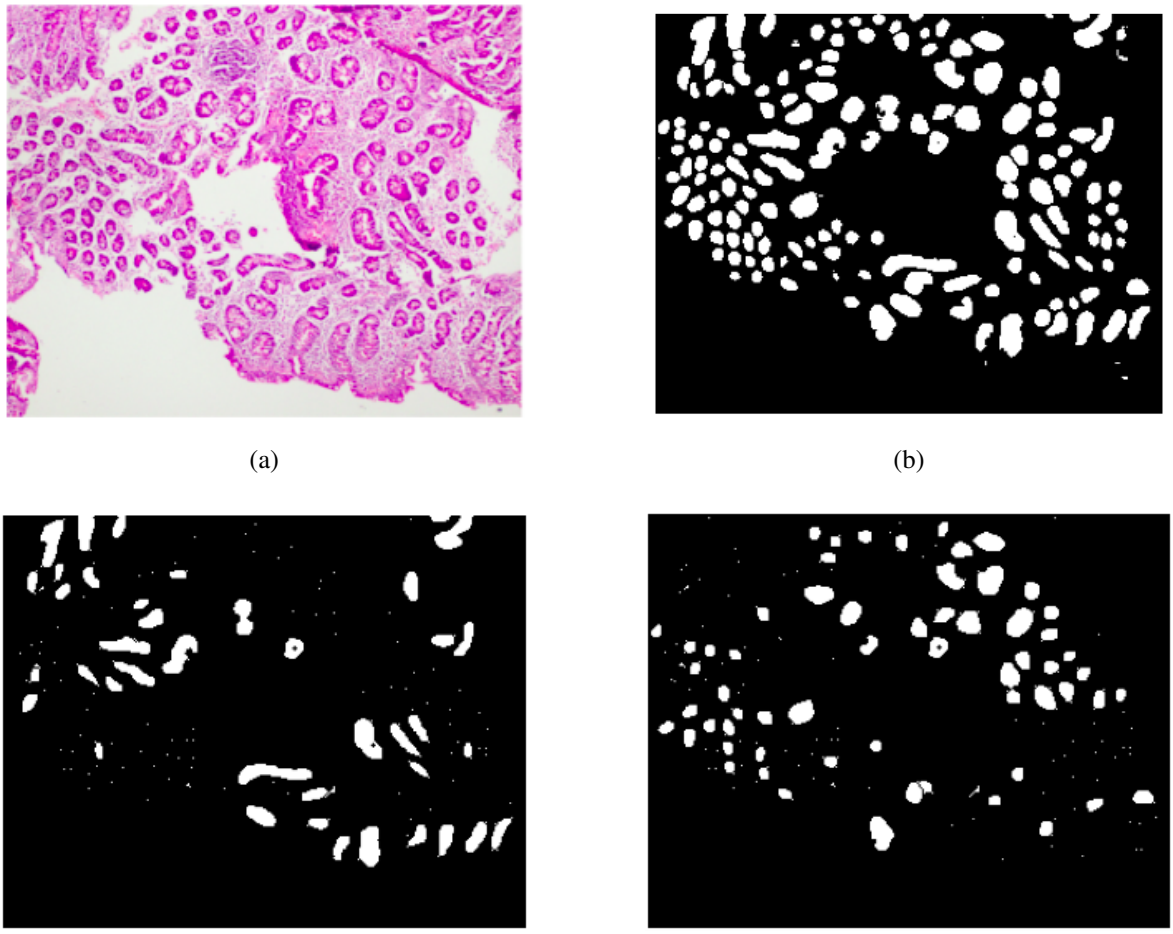
duodenal biopsies. One important morphological feature that is important to identify area of interpretability is the shape of crypts. The crypts are required to be elongated (test-tube rack like) in shape and they should be oriented towards the Good Villi for clinical interpretation. If a biopsy region contains round crypts, their interpretation is not possible. Figure 3.8 illustrates the differences in circular crypts and elongated crypts.



(a) (b)

Figure 3.7: (a) A biopsy image containing elongated crypts (Good Crypts) oriented towards Good Villi. (b) A biopsy image containing only Circular Crypts not fit for interpretation.

The segmentation stage in the cascaded model predicted Good Crypts with very less recall. Therefore, we combined the Good Crypts and Circular Crypts into a single class - Crypts which improved overall segmentation performance. The second stage in the cascaded model is exclusively to predict crypt edges as shown in Figure 3.6. These edge predictions are subtracted from the crypt segmentations from the first stage to obtain the inner pixels of crypts. The inner pixels are used initially to distinctly identify individual crypts using the



(a) (b)

(c) (d)

Figure 3.8: (a) A biopsy image containing two types of crypts. (b) Segmented crypts. (c) Good Crypts identified after circularity analysis. (d) Circular Crypts identified after circularity analysis.

Watershed Algorithm. Subsequently, ellipses are fit around each crypt and the eccentricities are measured using Eq 3.4

$$\text{Eccentricity}(e) = \frac{\text{Length of Minor Axis}}{\text{Length of Major Axis}} \tag{3.4}$$

We then define a threshold on e. If $0.5 < e < 1.0$, then the corresponding crypt is predicted as a Good Crypt. Otherwise, it is predicted as a Circular Crypt and therefore it is discarded from subsequent steps. Figure 3.7 illustrates the circularity analysis of crypts.

3.5 Discussion

Although the fully supervised methods make good predictions during inference time, training them is a lengthy procedure. The data annotation process is time-consuming and exhaustive in nature. Moreover during image segmentation, neither the cascaded system nor the Joint Learner explicitly learn representations of clinically important parameters like tissue morphology and the spatial relationships among different tissues. For a principled algorithm design that is clinically useful, targeted tissue representations must be learnt while minimizing the annotation effort. For this, we propose self-supervised methods in the succeeding chapters that make use of a larger unlabeled data corpus for learning targeted representations.

Chapter 4

Literature Survey

The distribution of real world data is extremely complicated. It is of paramount importance to separate the main factors of variation (called as the semantic structure) that are present in our data distribution in order to arrive at a meaningful understanding of each of its samples.

Since self-supervised learning exploits unlabelled data to learn image abstractions, a desirable property of any such method would be to meaningfully represent the semantic structure of the data they've been trained on. Popularly, there are two classes of self-supervised methods - the contrastive learning (CL) methods and the non-constrastive learning (NCL) methods.

The CL methods are essentially designed to perform instance-wise contrast leading to an embedding space where all instances are well-separated, and each instance is locally smooth (i.e. input with perturbations have similar representations). In other words, these methods encourage intra-class compactness and inter-class separability in the representation space. However, they suffer from a limitation: the representation is not encouraged to encode the semantic structure of data. This is because these methods consider two samples as a negative pairs as long as they are different instances in the training batch, regardless of whether or not they contain similar semantic structures. This is magnified by the fact that many negative samples are generated to form the contrastive loss which may contain similar semantic structures but are undesirably pushed apart in the representation space. Therefore, CL methods intuitively make most sense when individual data instances contain homogeneous semantics where all the structures in a particular data instance leads to a single semantic understanding

of that data instance. Since medical images often contain heterogeneous semantic structures in a single instance of data, using CL can be counter intuitive if our goal is to encode its semantics.

A recent work [29] tries to overcome this limitation of poor semantic encoding of CL methods using the so called *prototypes*. A prototype is defined as *"a representative embedding for a group of semantically similar instances"*. They replace the representation of the augmented view of data in the standard contrastive learning objective with this prototype. Thus, each data instance is assigned to a prototype while performing CL. Although, this well-designed heuristic provides a nice work-around for the limitations of conventional CL and does indeed try to encode semantics, it doesn't address the obvious question we asked at the beginning of this exposition relevant to the case of data in histopathology.

Another limitation of CL methods is studied as the problem of *class collison* where there is a possibility of contrasting two data instances actually belonging to the same class. [30] provides concrete theoretical bounds to determine the conditions under which empirical performance of CL methods will still be successful if sufficiently large negative instances of data are sampled during the training despite the possibility of class collision. It concludes *"the empirical minimizer of the unsupervised loss learned using sufficiently large number of samples will have good performance on supervised tasks"* which basically translates to, *"if we throw more data at the model, it will eventually learn."*

On the other hand, the NCL methods focus on training deep models to solve pretext tasks, which usually involve hiding certain information about the input and training the network to recover the missing information. In computer vision related applications, several researchers have proposed different self-supervised methods that focus on a variety of pretext tasks.

4.1 Self-Supervision by Solving Pretext Tasks

In computer vision related applications, several researchers have proposed different self-supervised methods that focus on a variety of pretext tasks. A pretext task is a puzzle which a model is asked to solve in order to capture representations that are naturally present in the data. In images, it maybe color, shapes, context etc that can be artificially changed before

asking the model to restore the images to their original state.

In [31], *surrogate classes* are constructed for training unsupervised models by applying random translations, scaling, rotations and altering the contrast of sub image patches containing objects or parts of objects. Given a set of transformed image patches of the same kind, a loss is minimized between the transformed sample and its surrogate class.

In [32], the authors have swapped image patch positions to break the image context and asked a deep CNN to reconstruct the original image. In the process, the model learns some general features of the whole images in the dataset.

In [33], the authors illustrate yet another pretext task of jigsaw puzzle reassembly. In this setting, given an image, the image patches are randomly jumbled and the model is asked to reassemble the patches in the right order.

In another work [34], the authors create a classification based pretext task of predicting one-of-N positions of an image patch given a reference image patch. In this, they choose a random patch in the image treating it as a pivot and sample one of 8 different target patches around it. These two patches are given as input to a deep CNN whose goal is to predict the position (in terms of class labels) of the target patch relative to the pivot patch.

In [35], the authors introduce yet another pretext task of image recolorization. Here, they feed the CNN with a grey-scale image and ask the model to predict colors in three different channels (R,G,B) at the decoder output. A vast variety of methods like the ones cited above focus on a single pretext task to learn self-supervised representations from images. While these standalone pretext task methods work very well, it is interesting to inquire into the viability of leveraging multiple pretext tasks together in a multi-task learning setting for self-supervision.

In [36], an SSL model is made to learn features by inpainting masked regions of the image. The masking happens on random image patches and the model is asked to paint inside the masked region by looking at the original image.

In one work as described in [37], the authors design an SSL model that has different decoder heads to solve multiple pretext tasks while having a common encoder. The idea is to leverage the combined power of representations learnt by solving multiple pretext tasks in a harmonised manner.

Since, these methods try to reconstruct contextual semantics in images, they are better suited to be applicable in our case when compared with CL methods where the notion of semantics is limited to semantics in natural images that are often easily discernible. Although, it must be noted that these methods explicitly or implicitly prescribe operations on image pixels and not on the image structure.

It is noteworthy that these SSL methods use large corpora of unlabelled data to learn strong representations by solving pretext tasks. Then, the learnt weights are finetuned on a small labelled data corpus for target tasks. It has been shown that these methods either beat state-of-the art fully supervised methods (trained on very large labelled datasets) or perform at par with them (while being fine-tuned on only a small portion of the labelled data).

The pretext tasks that various self-supervised methods present learn generic image representations. In the medical imaging setup with rich context information stored in various tissues in terms of colour, morphology, texture and the relationships among different tissues in the image, it is hard to explicitly say exactly what representations these generic pretext tasks learn. In that, the generic pretext tasks such as jigsaw, relative patch prediction by sampling random pivot patches etc learn poor representations. In other words, it is virtually impossible to verify if the model has properly learnt to represent a certain kind of organ or a tissue having clinical importance.

The SSL methods that solve generic pretext tasks cannot guarantee that they learn only important representations. For example, if we were to apply the image context restoration as a pretext task in the case of duodenal histology images by randomly sampling two patches from the background of the image and swapping their positions, the model will learn representations that are of no clinical significance. Similarly, if we perform the image inpainting pretext (as in [36]) task by randomly masking sub patches in the background, a similar problem may occur.

More generally, in any medical imaging setup, a pretext task that is applied on regions of no clinical significance like background or tissues that aren't considered in the process of diagnosis (Denudated Villi in the case of duodenal biopsies or regions like the neck, shoulders that are partially visible in chest X-rays.) will lead to learning unwanted representations.

In such cases, a better practice would be to apply pretext tasks specifically on regions

of diagnostic importance. That would be to swap positions of sub patches (in the context-resotoration pretext task) or mask only those sub patches (in the image inpainting pretext task) that belong to a diagnostically relevant region.

Although, it is also worth noting that not all pretext tasks need guidance. For solving the image recolorization task, it is better to convert the whole image into gray-scale rather than just having small regions converted into gray for learning stronger colour representations. Therefore, in this work, we describe two clinically guided pretext tasks. We conduct two kinds of experiments. One method solves such a pretext task with deep feature reconstruction and the other task involves reconstruction of image superpixels.

4.2 Prior Work

Since we are working with a completely novel and unique dataset, no direct prior work has been conducted on it. However, [38] tackled the problem of predicting Celiac Disease by proposing an end-to-end classifier. Given a histological slide, the algorithm classifies it as either having Celiac Disease or non-specific duodentitis or simply classifies it as normal tissue. Our methods introduce clinical knowledge into the learning process in the form of pretext tasks and requires limited labelled data. A more recent work [39] proposes a machine learning approach to detect Celiac disease based on the Marsh scoring system.

In contrast, we emphasize the clinical relevance of morphological patterns seen in histopathological slides. We develop self-supervised methods as opposed to fully supervised methods in literature. Our proposed method attempts to incorporate learning of these morphologies during the pre-training stage.

Chapter 5

Self Supervised Learning methods

5.1 Introduction

In this chapter, three self-supervised learning methods are described. The first method uses a popular unsupervised image segmentation algorithm that groups perceptually similar pixels with each other and then masks them. Then, a neural network is asked to reconstruct the noisy images thereby learning strong image representations.

The second method uses a very weak fully supervised model (mentor network) trained on a very small number of labelled images and guides a self-supervised model (mentee network) to learn stronger representations on a corpus of a larger unlabelled data by solving clinically motivated pretext tasks on the duodenal biopsy dataset.

For both the methods, only unlabelled images are used during self-supervised pretraining. Later, a small subset of labelled images is used for finetuning the semantic segmentation task. However, for the second method - the same subset of labelled images is used for training the weak mentor network.

Evidently, the role of unlabelled data is important for the success of these methods. Once we have sufficient data to work with, the challenge then is to design the right pretext tasks. But exactly how large should the unlabelled data corpus be is an open-ended question. The findings in [40] show that even a single image suffices, with self-supervision and data augmentation, to learn the first few layers of deep neural networks as well as using millions of images and full supervision. For the deeper layers of the networks, its concluded that

self-supervision remains inferior to strongly supervised methods even if millions of images are used for training them. The authors also conclude that adding more unlabelled data is unlikely to improve the performance of these models. For our experiments, we have used a little over 1000 unlabelled images for pretraining.

5.2 Super Pixel Inpainting as a Pretext Task

A Super Pixel can be defined as a group of image pixels that share common perceptual characteristics (like color intensity). Super Pixels provide more information about a region inside an image as compared to pixels. They align better with edges in the image when compared to simple rectangular image patches.

SLIC [41] is an unsupervised algorithm that uses the k-means clustering to give Super Pixels. We use the SLIC algorithm to first identify Super Pixels in an image from the unlabelled dataset D_U. Figure 5.1 is an illustration of the training procedure for this method. The Super Pixels inside the biopsy region are randomly masked. Figure 5.2 shows the masked Super Pixels against the corresponding real histological image.

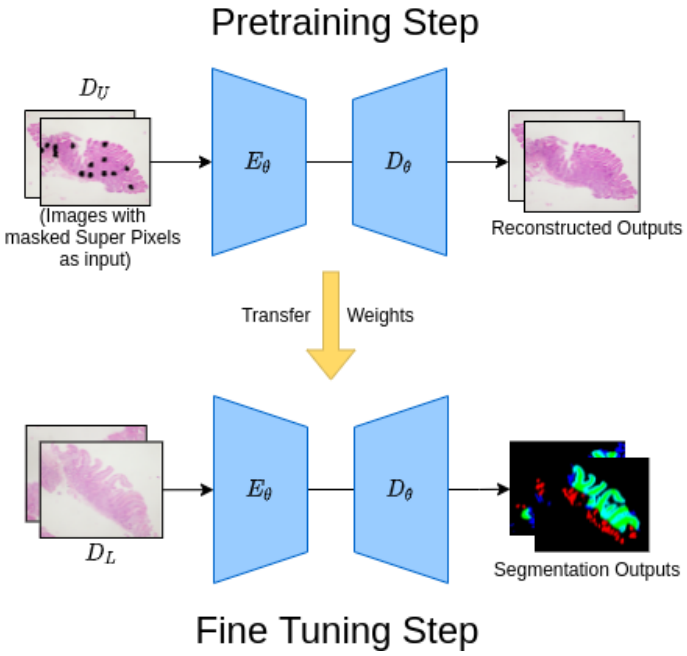


Figure 5.1: Super Pixel Inpainting as a self-supervision pretext task.

The masked image is then given to a neural network for reconstruction. We use the structural similarity [42] loss function during the pretraining step. Figure 5.3 shows the reconstructed output corresponding to the masked input after the pretraining step. After pretraining, the model weights are finetuned on the segmentation task using a small labelled dataset D_L. The idea is to learn the representations of tissues and other structures inside the biopsy area that maybe useful for semantic segmentation and subsequently to identify Interpretable Regions.

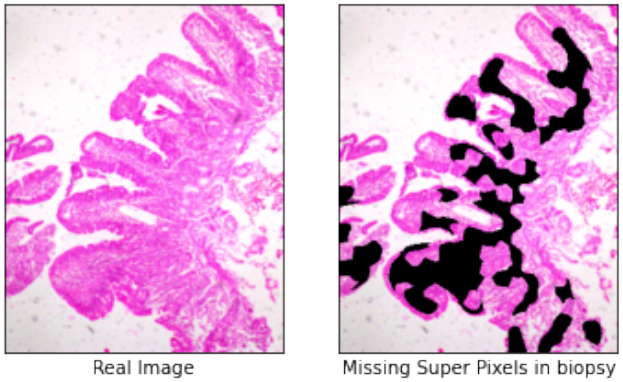


Figure 5.2: Masking of Super Pixels from a histological image. The masked image is fed to a neural network for reconstruction during self-supervised pretraining.

Figure 5.3: The masked super pixels, the reconstructed output and the real image.

5.3 Super pixel inpainting with morphology restoration

There are two limitations to the above method. In that, it isn't designed to consider morphological features while doing image reconstruction. The other limitation is that, the inpainting simply learns to fill the masked out pixels with a shade of pink, thus leading to a degenerate solution. To focus on morphology and to avoid the degenerate condition, we propose a complementary task that distorts the shapes of tissues using the method described in [43] and the autoencoder is tasked with restoring the morphology of various tissues in the image.

In our experiments for this method, we use the SLIC algorithm to mask random superpixels from the H&E image and apply the elastic distortion on top of it. Figure 5.4 shows the composite distortions on an HE image which is treated as an input at the pre-training step in figure 5.1. The subsequent steps remain the same. We obtain an improvement in performance with these modifications in our training mechanism. Table 5.1 shows the segmentation performance with these changes. The implementation of this method can be found here.

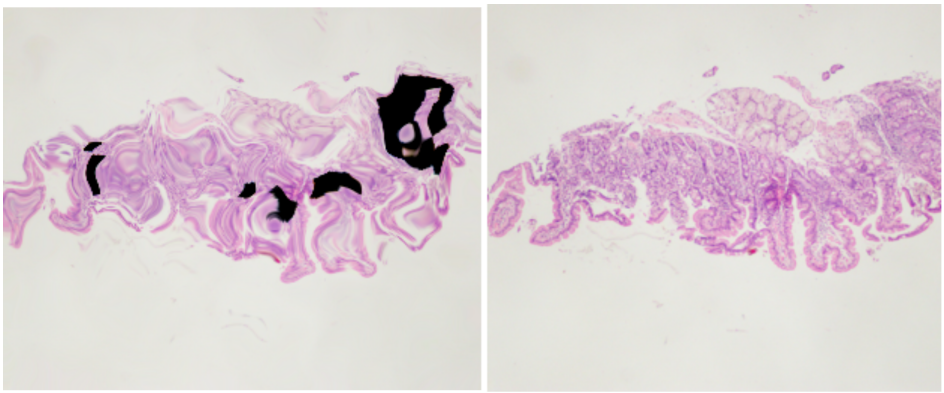


Figure 5.4: Composite distortion: Masking of super pixels along with morphology deformation of a histological image. These images are fed to a neural network for reconstruction during self-supervised pretraining.

Tissue	Finetuned (Proposed method)	Full Supervision
Good Crypts	0.599	0.453
Good Villi	0.515	0.484
Epithelium	0.561	0.425
Average	0.558	0.454

Table 5.1: The performance comparison of the super pixel method with the fully supervised method. 50 labelled images were used to train the supervised network. The same images were used for finetuning. 1150 unlabelled images were used for self-supervision. Reported metric is the Dice Score.

5.4 Deep Feature Reconstruction for Representation Learning of Tissue Morphology

Despite better performance of the Super Pixel method, we wish to gain fine-grained control over what representations are exactly learnt. We design a novel pretext task for this purpose called the Elastic Deformation and Masking (EDM). In that, this pretext task precisely targets the classes of interest during the self-supervised pretraining stage. To do this, initially, we train a fully supervised semantic segmentation model on a small labelled dataset D_L. This model is used for segmenting tissues from the unlabelled dataset D_U. The outputs thus obtained are treated as pseudolabels. Although noisy, they are used to randomly mask some of segmented tissues (refer Figure 5.5).

The motivation is to specifically teach the model to learn the representations of tissues like Good Crypts, Good Villi and Epithelium. Randomly masking some of these objects that we get from the fully-supervised outputs motivates the model to do an inpainting task only inside the specific tissue regions while elastic deformation will teach the model morphological representations of these tissues.

Deep Feature Reconstruction: During Pretraining we motivate the self-supervised model to partially mimic the latent features from the mentor network. Figure 5.6 shows the implementation idea. L_{df} is MSE loss between the deep features of the supervised model and the corresponding features of the SSL model. L_{SSIM} is the structural similarity loss aplied be-

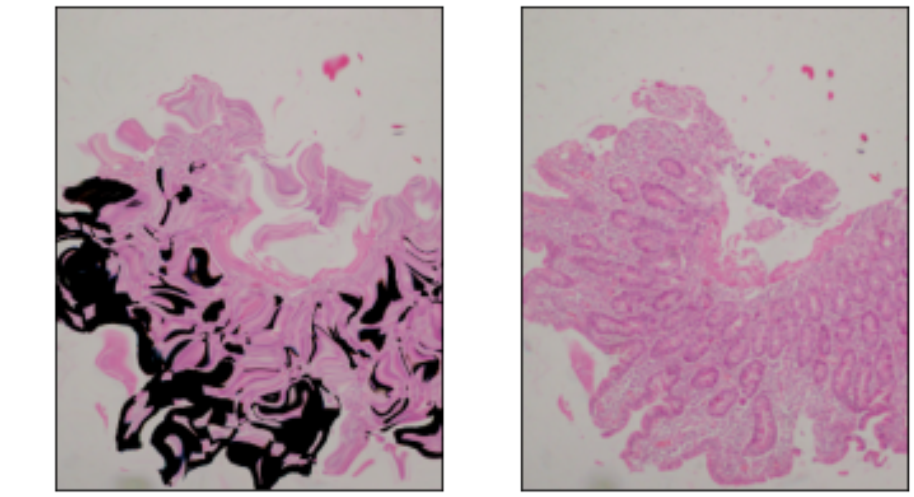


Figure 5.5: The EDM pretext task. Good Villi, Good Crypts and Epithelium are masked and deformed.

tween reconstructed image and the original input. The overall loss used during the training procedure is given by:

$$L = L_{SSIM} + \lambda L_{df} \tag{5.1}$$

Where λ is a hyperparameter set to 0.1. Table 5.2 shows the comparison of this method with its fully supervised counterpart. Figure 5.7 shows the reconstructed image from the model. It should be noted that the tissue morphologies have been reconstructed faithfully by the proposed SSL model. These experiments are implemented here.

5.5 Implementation Details

For both the experiments, the backbone used is an Attention-Unet. The learning rate decreases from 10E-3 through 10E-6 with a step size of 0.1. The patience is set to 20. Training stops if no improvement happens after 35 epochs. This setting is kept the same for both finetuning as well as pretraining with unlabelled images. All experiments were conducted on an NVIDIA V100 with 32GB RAM. We use 50 labelled images for finetuning, 40 labelled images as validation images and a separate 60 labelled images for model evaluation.

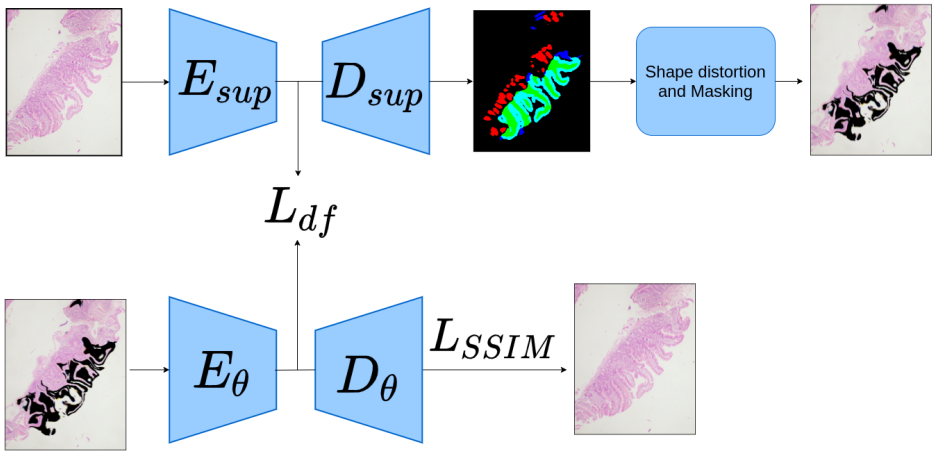


Figure 5.6: The elastic deformation and masking pretext task. Good Villi, Good Crypts and Epithelium are masked and deformed.

Method	Data Subset	Crypts	Villi	Epithelium	B. Gland	Average
Full Supervision	50 D_L	0.43	0.53	0.49	0.63	0.52
EDM	50 D_L + 1150 D_U	0.47	0.46	0.56	0.82	0.58
Full Supervision	300 labelled	0.68	0.6	0.64	0.85	0.69
EDM	50 D_L + 1150 D_U	0.68	0.64	0.67	0.84	0.71

Table 5.2: Performance comparison of the EDM method with fully supervised counterpart. 50 labelled images were used to train the supervised network. The same images were used for finetuning. 1150 unlabelled images were used for self-supervision. Reported metric is the Dice Score.

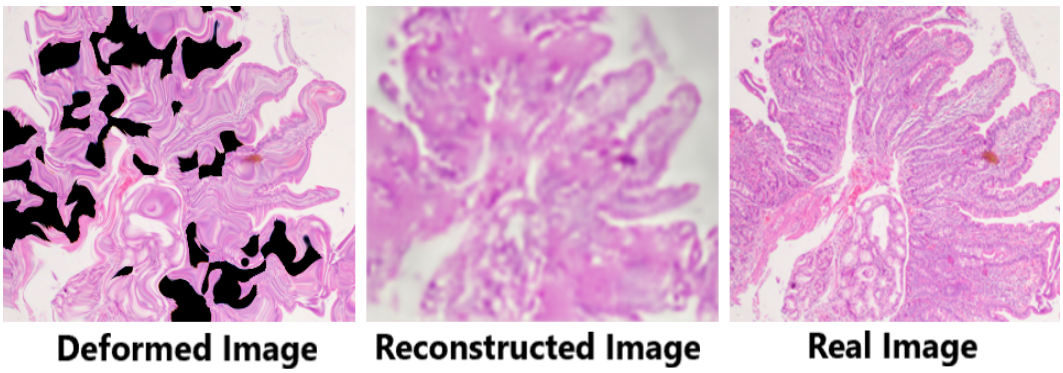


Figure 5.7: Tissue inpainting and Morphology restoration as a pretext-task

Chapter 6

Conclusion

6.1 Summary

A novel richly annotated dataset of the human duodenum is introduced. Two fully supervised methods are described that perform localization of areas of interpretability. Two self-supervised methods for learning meaningful representations of diagnostically important regions in medical images are described. Two clinically motivated pretext tasks are introduced which are shown to learn robust representations of specific tissue regions from the histological images. With superior performance as compared to corresponding supervised baselines, the study establishes some important directions which can be explored while developing self-supervised learning methods for general medical image analysis using deep learning.

6.2 Future Work

6.2.1 Villi Lengths Measurement

For the Q-histological criteria, measuring of Villi lengths is important. Different heuristics can be applied to measure the individual Villi Lengths but a neural method works the best. We annotated 25 images demarcating the length along the Villous fingers. Then, this set is randomly split into 20 train and 5 validation samples. A segmentation model is trained to predict different strands of measurements that can be isolated as connected components and

measured individually. Focal Tversky loss is used to handle the thin segments that implicitly cause class imbalance. We achieve a dice score of 53% over Villi measurements given the Villi mask priors. Future work is motivated in this direction where these outputs could be post processed to obtain accurate measurements of individual Villous Lengths. Along with this, the measurement of Crypt depths should be performed for a complete software solution to automatically grade Celiac Disease. The implementation for Villi lengths measurement can be found here.

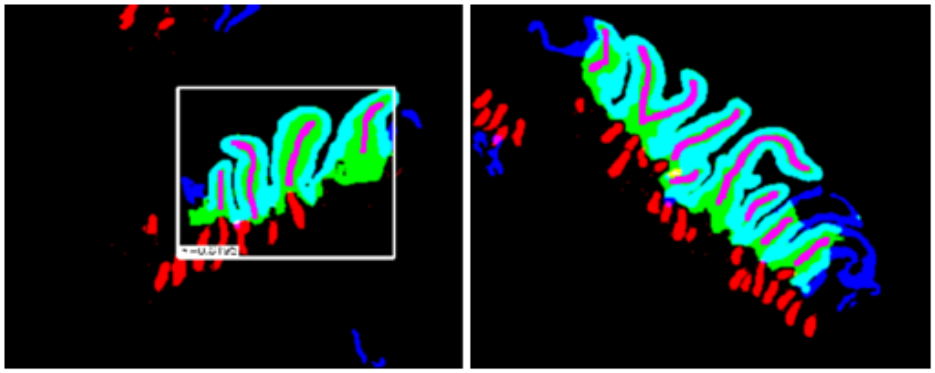


Figure 6.1: Sample predictions by the segmentations for demarcating the Villi lengths.

6.2.2 Counting Intra-Epithelial Lymphocytes

Counting the Intra-Epithelial Lymphocytes (IELs) in the epithelial region of the Villi is another important quantitative parameter in the Q-histology classification system. We exhaustively annotated 65 images for IELs. Out of these, we used 45 images for training and 8 were used for validation. Various data augmentation techniques were done to compensate for the limited data. This problem can be formulated as an object detection problem. However, the IELs are extremely small when compared to the size of the image.

To circumvent this issue and retain global context at the same time, we use EfficientDet for localization. Since the Epithelial region on the periphery of the Villi are important, we use a segmentation network to first segment out this region. Once that's done, we mask the whole image and retain only the Epithelial Belts during training. Of the visible pixels, we make patches containing IELs and these patches are given to the object detection model to localise the tiny cells. Figure 6.2 illustrates this process visually. The mAP score we obtain with this

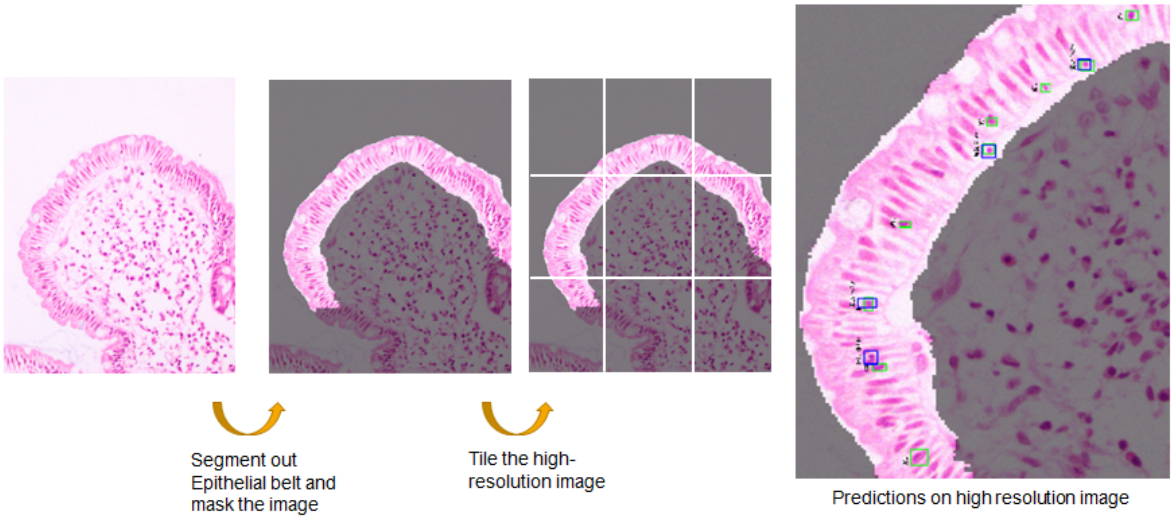


Figure 6.2: Cascading Networks for detecting IELs in high resolution.

method is 0.34. We find that this score is significantly boosted (0.54) when the ground truth epithelial belts are supplied to the network during training. Hence, we motivate future work to solve the bottleneck of accurately segmenting the epithelial belts at the periphery of the Villi. Find the implementation of IEL counting here.

Bibliography

[1] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[8] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, *et al.*, "The open images dataset v4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[9] A. Ess, T. Müller, H. Grabner, and L. Van Gool, "Segmentation-based urban traffic scene understanding.," in *BMVC*, vol. 1, p. 2, Citeseer, 2009.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[12] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.

[13] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 24–32, 2015.

[14] A. Fasano and C. Catassi, "Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum," *Gastroenterology*, vol. 120, no. 3, pp. 636–651, 2001.

[15] A. Fasano, I. Berti, T. Gerarduzzi, T. Not, R. B. Colletti, S. Drago, Y. Elitsur, P. H. Green, S. Guandalini, I. D. Hill, *et al.*, "Prevalence of celiac disease in at-risk and not-at-risk groups in the united states: a large multicenter study," *Archives of internal medicine*, vol. 163, no. 3, pp. 286–292, 2003.

[16] J. C. Gomez, G. S. Selvaggio, M. Viola, B. Pizarro, G. La Motta, S. De Barrio, R. Castelletto, R. Echeverrıa, E. Sugai, H. Vazquez, *et al.*, "Prevalence of celiac disease in argentina: screening of an adult population in the la plata area," *The American journal of gastroenterology*, vol. 96, no. 9, pp. 2700–2704, 2001.

[17] M. Rewers, "Epidemiology of celiac disease: what are the prevalence, incidence, and progression of celiac disease?," *Gastroenterology*, vol. 128, no. 4, pp. S47–S51, 2005.

[18] G. K. Makharia, A. K. Verma, R. Amarchand, S. Bhatnagar, P. Das, A. Goswami, V. Bhatia, V. Ahuja, S. Datta Gupta, and K. Anand, "Prevalence of celiac disease in the northern part of india: a community based study," *Journal of gastroenterology and hepatology*, vol. 26, no. 5, pp. 894–900, 2011.

[19] P. Das, G. P. Gahlot, A. Singh, V. Baloda, R. Rawat, A. K. Verma, G. Khanna, M. Roy, A. George, A. Singh, and et al., "Quantitative histology-based classification system for assessment of the intestinal mucosal histological changes in patients with celiac disease," *Intestinal Research*, vol. 17, no. 3, p. 387–397, 2019.

[20] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011.

[21] L. Fulford, D. Easton, J. Reis-Filho, A. Sofronis, C. Gillett, S. Lakhani, and A. Hanby, "Specific morphological features predictive for the basal phenotype in grade 3 invasive ductal carcinoma of breast," *Histopathology*, vol. 49, no. 1, pp. 22–34, 2006.

[22] A. Katayama, M. S. Toss, M. Parkin, T. Sano, T. Oyama, C. M. Quinn, I. O. Ellis, and E. A. Rakha, "Nuclear morphology in breast lesions: refining its assessment to improve diagnostic concordance," *Histopathology*, vol. 80, no. 3, pp. 515–528, 2022.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[24] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," 2019.

[25] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[27] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[28] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," 2020.

[29] J. Li, P. Zhou, C. Xiong, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*, 2020.

[30] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*, 2019.

[31] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015.

[32] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.

[33] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, pp. 69–84, Springer, 2016.

[34] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

[35] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*, pp. 649–666, Springer, 2016.

[36] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

[37] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

[38] J. W. Wei, J. W. Wei, C. R. Jackson, B. Ren, A. A. Suriawinata, and S. Hassanpour, "Automated detection of celiac disease on duodenal biopsy slides: A deep learning approach," *Journal of pathology informatics*, vol. 10, 2019.

[39] J. E. W. Koh, S. De Michele, V. K. Sudarshan, V. Jahmunah, E. J. Ciaccio, C. P. Ooi, R. Gururajan, R. Gururajan, S. L. Oh, S. K. Lewis, P. H. Green, G. Bhagat, and U. R. Acharya, "Automated interpretation of biopsy images for the detection of celiac disease using a machine learning approach," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106010, 2021.

[40] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," *arXiv preprint arXiv:1904.13132*, 2019.

[41] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[43] P. Y. Simard, D. Steinkraus, J. C. Platt, *et al.*, "Best practices for convolutional neural networks applied to visual document analysis.," in *Icdar*, vol. 3, Edinburgh, 2003.