BOOSTING FULLY CONVOLUTIONAL NETWORKS FOR GLAND INSTANCE SEGMENTATION IN HISTOPATHOLOGICAL IMAGES

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Gözde Nur Güneşli

August 2019

Boosting Fully Convolutional Networks for Gland Instance Segmentation in Histopathological Images
By Gözde Nur Güneşli
August 2019

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Çiğdem Gündüz Demir(Advisor)

Shervin Rahimzadeh Arashloo

Ramazan Gökberk Cinbiş

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

BOOSTING FULLY CONVOLUTIONAL NETWORKS FOR GLAND INSTANCE SEGMENTATION IN HISTOPATHOLOGICAL IMAGES

Gözde Nur Güneşli
M.S. in Computer Engineering
Advisor: Çiğdem Gündüz Demir
August 2019

In the current literature, fully convolutional neural networks (FCNs) are the most preferred architectures for dense prediction tasks, including gland segmentation. However, a significant challenge is to adequately train these networks to correctly predict pixels that are hard-to-learn. Without additional strategies developed for this purpose, networks tend to learn poor generalizations of the dataset since the loss functions of the networks during training may be dominated by the most common and easy-to-learn pixels in the dataset. A typical example of this is the border separation problem in the gland instance segmentation task. Glands can be very close to each other, and since the border regions contain relatively few pixels, it is more difficult to learn these regions and separate gland instances. As this separation is essential for the gland instance segmentation task, this situation arises major drawbacks on the results. To address this border separation problem, it has been proposed to increase the given attention to border pixels during network training either by increasing the relative loss contribution of these pixels or by adding border detection as an additional task to the architecture. Although these techniques may help better separate gland borders, there may exist other types of hard-to-learn pixels (and thus, other mistake types), mostly related to noise and artifacts in the images. Yet, explicitly adjusting the appropriate attention to train the networks against every type of mistake is not feasible. Motivated by this, as a more effective solution, this thesis proposes an iterative attention learning model based on adaptive boosting. The proposed *AttentionBoost* model is a multi-stage dense segmentation network trained directly on image data without making any prior assumption. During the end-to-end training of this network, each stage adjusts the importance of each pixel-wise prediction for each image depending on the errors of the previous stages. This way, each stage learns the task with different attention forcing the stage to learn the mistakes of the earlier

stages. With experiments on the gland instance segmentation task, we demonstrate that our model achieves better segmentation results than the approaches in the literature.

Keywords: Deep learning, attention learning, adaptive boosting, gland segmentation, medical image segmentation.

ÖZET

HISTOPATOLOJIK GÖRÜNTÜLERDE BEZ ÖRNEğI BÖLÜTLEMESI ÇIN TAM EVRIşIMSEL Ağ GÜÇLENDIRMESI

Gözde Nur Güneşli
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Çiğdem Gündüz Demir
Ağustos 2019

Mevcut literatürde, tam evrişimsel sinir ağları (FCN'ler), bez bölütleme de dahil olmak üzere yoğun tahmin işleri için en çok tercih edilen mimarilerdir. Öte yandan, öğrenmesi zor pikselleri doğru şekilde tahmin etmek için bu ağları yeterince eğitmek önemli bir zorluktur. Bu amaçla geliştirilmiş ek stratejiler olmadan, ağlar veri setinin zayıf genellemelerini öğrenmeye meyillidir. Buna neden eğitim sırasında ağların kayıp fonksiyonlarına veri setindeki en yaygın ve öğrenmesi kolay piksellerin hakim olabilmesidir. Bez örneği bölütleme işindeki sınır ayrımı problemi bu duruma tipik bir örnektir. Bezler birbirine çok yakın olabilir ve sınır bölgeleri nispeten az piksel içerdiğinden, bu bölgeleri öğrenmek ve bez örneklerini ayırmak daha zordur. Bu ayırma, bez örneği bölütleme işi için esas olduğundan; bu durum, sonuçlarda büyük dezavantajlara yol açar. Bahsedilen sınır ayırma problemi için, bu piksellerin bağıl kayıp katkısını artırarak veya sınır algılamayı mimariye ek bir görev olarak ekleyerek ağ eğitimi sırasında sınır piksellerine verilen dikkatin artırılması önerilmiştir. Her ne kadar bu teknikler bezlerin sınırlarını daha iyi ayırmaya yardımcı olsa da, görüntülerde çoğunlukla gürültü ve artifaktlara bağlı başka öğrenmesi zor piksel türleri (ve bundan dolayı başka hata türleri) olabilir. Ancak, ağların her türlü hataya karşı eğitilmesi için uygun dikkatin açıkça ayarlanması mümkün değildir. Bunu motivasyonla, daha etkili bir çözüm olarak, bu tez uyarlamalı güçlendirmeye dayanan yinelemeli bir dikkat öğrenme modeli önermektedir. Önerilen bu *AttentionBoost* modeli; önceden bir varsayımda bulunulmadan, doğrudan görüntü verileri üzerinde eğitilen çok aşamalı bir yoğun tahmin ağıdır. Bu ağın uçtan uca eğitimi sırasında, her aşama, önceki aşamaların hatalarına bağlı olarak, her görüntüdeki her piksel için tahminin önemini ayarlar. Bu şekilde, her aşama, kendisini önceki aşamaların hatalarını öğrenmeye zorlayan farklı bir dikkatle ilgili işi öğrenir. Bez örneği bölütleme

işi üzerinde yapılan deneyler, modelimizin literatürdeki yaklaşımlardan daha iyi sonuçlar elde edebildiğini göstermiştir.

Anahtar sözcükler: Derin öğrenme, dikkat öğrenme, uyarlanabilir güçlendirme, bez bölütleme, medikal görüntü bölütlemesi.

Acknowledgement

I wish to thank various people for their contributions to my studies:

Foremost, I would like to express my deepest appreciation to my advisor Assoc. Prof. Dr. Çiğdem Gündüz Demir for everything. It is a great honor for me to do my M.Sc. studies under her supervision. Her guidance and expertise made this an inspiring experience for me.

Also, I am extremely grateful to my jury members Asst. Prof. Shervin Rahimzadeh Arashloo and Asst. Prof. Gökberk Cinbiş for reviewing and commenting on this thesis.

I would like to thank to the Scientific and Technological Research Council of Turkey (TÜBİTAK) for providing financial assistance during my study, through the project TÜBİTAK 116E075.

I am deeply indebted to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. I am also very grateful to my sister and my friends who have supported me along the way. Without them, this accomplishment would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diagnosis and grading of many neoplastic diseases including cancer is based on microscopic analysis of sections of biopsies and tissue specimens, examined by pathologists. Because of the increasing number of cancer patients, time required for the examination process, inter- and intra-observer variability among the pathologists, it is worthwhile to develop computer based methods that automate this process [1].

1.1 Motivation

With the advance of deep learning on image related tasks, many methods to analyze medical images have constructed their models based on convolutional neural networks (CNNs), especially using fully convolutional networks (FCNs) for the segmentation tasks [2]. Although these deep learning approaches have provided significant improvements over the traditional methods, there are still some challenges that make the segmentation task difficult to automate for glandular structures in histopathological images. The most significant challenge is the non-homogeneity of the appearances of these structures. To elaborate, there are variations in gland appearances, moreover, this variation increases with the

existence of cancer and irregularity becomes more apparent with the increasing cancer grade. Additionally, the tissue sectioning and staining processes can cause differences such as deformations, color differences, and artifacts on the image. When these challenges combine with the limitations of the annotated data available to train a supervised segmentation model, the model tends to overfit to a local minimum generalization of the training set easily. This generalization is likely to be affected by majority classes and easy-to-learn in-class appearances. As a result, the models become prone to poor generalizations, yielding low accuracy for pixels of minority classes as well as for hard-to-learn pixels.

The most common approach to solve this problem is to adjust the loss weights, which determine the effect of each pixel prediction to the total loss function in the training phase. Commonly, as a solution to the class-imbalance problem, many studies train their networks using pre-computed loss weights that increase the relative weight of minority class predictions in the loss function. Although with this approach the network can be forced to give more attention to learning the minority class, this approach does not solve other imbalance issues arising from the nature of the glandular structures and the preparation process of the histopathological images. Since instances of a particular class usually have different appearances with different frequencies, giving the same coefficient to all predictions of the same class usually results in poor learning of less frequent or hard-to-learn parts in this class. For example, being able to separate touching components accurately is a necessity for the instance segmentation task. However, this requires accurate predictions of border pixels and since these border pixels usually have low frequencies in the class that they belong to, they are harder to learn.

To overcome this "border separation problem", one proposed solution is to increase the given attention to the classification of the border pixels. The U-net model by Ronneberger et al. [3] proposes to solve this problem using pre-computed weight maps for the loss function obtained from a function dependent on the distances to the nearest objects from each pixel for each training image. As another solution, the DCAN model by Chen et al. [4] attempts to solve the same problem with a multi-task architecture, in which border prediction is considered as an

additional task to the main task of segmentation and learned with shared features. Then, they combine the predicted border map with the segmentation map to obtain the final predictions. Xu et al. [5] expand this idea adding a detection task (bounding box information) to the multi-task architecture. Although all of these solutions may help alleviate the mistakes related to incorrect prediction of border pixels, there may exist other hard-to-learn pixels, which cause different types of mistakes due to the nature of the problem at hand (see Figure 1.1). Since these solutions define their attentions externally and manually, to be robust against multiple mistake types, they need to define different loss weights or new additional tasks with respect to each different mistake type. Since adjusting different hand-crafted weights or additional multi-task learners for every hard-to-learn appearance with different characteristics is a very challenging task, a method to automatically learn the convenient attention on the training images could be very beneficial.

Besides using predefined attentions, multi-stage models have been proposed to improve the predictions of the FCN models. These multi-stage models are combinations of multiple iterative networks, each of which learns to refine the prediction map of the previous network(s) by taking the image and the previous prediction map as inputs [6, 7, 8, 9, 10]. After a certain number of iterations, they use the last prediction map as the final segmentation. In these models, each network in each stage learns the same task with the same objective, and the networks in consecutive stages are expected to learn to refine the errors of the previous stages implicitly. However, since the objective, and thus, the attention of all these stages are same, they will still suffer from the influence of the aforementioned problems.

1.2 Contribution

In this thesis, we propose an iterative attention learning framework based on adaptive boosting for the effective and robust segmentation of glandular structures in histopathological images. This framework, which we call *AttentionBoost*,

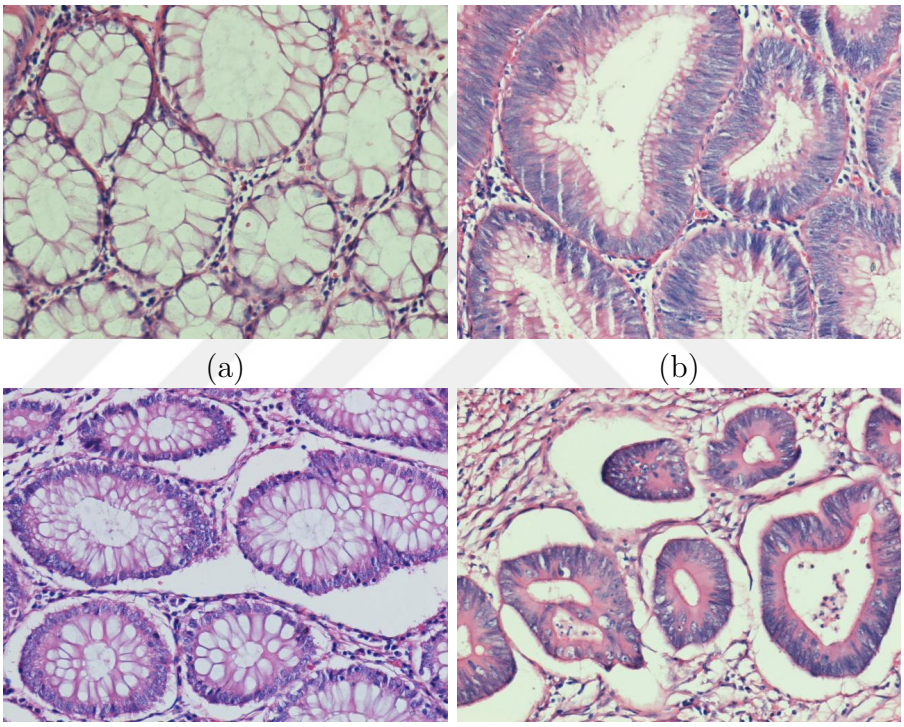


Figure 1.1: Examples of histopathological images of colon glands. The images shown in (a) and (b) illustrate the cases in which the glands are very close to each other. For these cases, it is more difficult to correctly classify the boundary pixels. Additionally, histopathological images typically contain noise and artifacts due to the tissue preparation procedures. The images given in (c) and (d) contain such kind of artifacts. It is common for gland segmentation algorithms to identify some of these large white artifacts as false glands. These are the images consisting of (a)-(c) normal glands and (b)-(d) cancerous glands.

is constructed as a multi-stage system that contains a fully convolutional segmentation network in each stage. By introducing a new loss adjustment method for a dense prediction model, the segmentation (sub)networks of each stage of this system is forced to have a specific attention to decrease the errors of the previous stages. This proposed loss adjustment method is inspired by the Adaboost algorithm [11]. It proposes to modulate the attention of each segmentation network during training, adjusting the relative contribution of each pixel prediction to the loss function of each network while the network weights are learned at the same time. Then, in the testing phase, the intermediate results of the networks of all stages are combined for the final predictions. Our experiments demonstrate that our model leads to superior test results on the gland instance segmentation task compared to the existing approaches in the literature. This is attributed to the fact that the proposed model not only pays attention to border pixels but to other hard-to-learn pixels as well, which are mostly related to noise and artifacts in the images.

The proposed *AttentionBoost* model is different than the previous studies in the following aspects. In the literature, there are FCN based models that define their attention before training their networks [3, 4, 12], commonly to solve the border separation problem. While those models have a single attention, which is predefined externally and manually, the *AttentionBoost* model adjusts its attention at each stage automatically depending on the errors of the previous stages. Thus, it does not require any predefined attention.

AttentionBoost is also different than the multi-stage models [6, 7, 8, 9, 10] in the literature. In these models, each network in each stage learns the same task with the same objective without changing its attention. Although the *AttentionBoost* model uses a multi-stage architecture similar to these models, because of its proposed loss adjustment method, it is able to regulate the attention of each network to a different aspect of the objective.

In the literature, there also exist studies that use predefined weights for the objective functions to solve the class-imbalance problem. They use a constant weight for all predictions of the pixels in the same class [13, 14, 15]. Those weights

are selected to increase the given attention to the minority class. Different than these studies, the proposed *AttentionBoost* model adjusts the weights in the loss function allowing different weights for the pixel-wise predictions of the same class.

There is only one study that attempts to learn the loss weights for the object detection task on image data. However, different than our model, this study does not construct or iteratively train multiple networks, but instead focuses on the training of a single stage network [16]. The loss weight for each individual object is updated at each epoch during the training and the next epoch uses the same updated weight in every pixel within the same object bounding box. This approach might increase the importance of misdetected and more difficult objects for learning in later epochs. However, because it uses a single network, the common type of incorrectly/correctly detected objects may dominate the loss function, which makes it difficult to explicitly concentrate on several detection subtasks with different levels of difficulty at the same time. On the other hand, the *AttentionBoost* model is constructed with multiple networks, each of which can have a different attention. This enables each stage to better concentrate on a different aspect of the task. In addition, this previous study [16] uses the same loss weight for all pixels of the same object (bounding box), without any consideration of being given to their pixel-wise contributions. By contrast, depending on the difficulty of learning the pixels, *AttentionBoost* updates the loss weight for each pixel individually.

In the literature, there are also studies that use the Adaboost algorithm [11] with a neural network architecture [17, 18, 19, 20, 21]. Yet, these studies do not include a dense prediction task using an FCN, but rather they are designed for the classification of an image. Thus, for each image, they use the same attention either by arranging different training sets for each learner or by arranging loss weights for each learner's training instances (images). On the other hand, *AttentionBoost* uses the idea to adjust pixel-wise loss weights of a dense segmentation model. These non-dense models, intended for the task of labelling an entire image with a single class, are outside the scope of this thesis.

1.3 Outline

This thesis is organized as follows. Chapter 2 gives a summary of the related work in the literature along with the background information about the problem domain and the deep learning methods. Chapter 3 presents our methodology including the details of the proposed loss adjustment method and the framework architecture. Chapter 4 provides the experimental settings, including the dataset, metrics, and comparison methods used for evaluation. Then, Chapter 5 reports the test results and gives a discussion about the results. Finally, Chapter 6 contains the conclusion remarks and future aspects of this thesis.

Chapter 2

Background

The main purposes of histopathological image analysis are to determine the disease and its state and progression accurately based on digitized histology slides. The tasks in the current literature can be divided into three main groups, all serving for these purposes: segmentation (e.g., segmenting glands), detection (e.g., counting cells), and classification (e.g. differentiating benign and malignant structures). Segmentation, which is also the topic of this thesis, is a fundamental step in the automated diagnosis process of many neoplastic diseases including colon adenocarcinoma. Its role is to identify the location of relevant areas (i.e., glands) in an image. Identification of these areas requires a reliable segmentation tool.

In this chapter, we first give a description of the domain for the gland instance segmentation task. Specifically, we will give information about colon adenocarcinoma, the relationship of this disease with the gland structures and the importance of the instance segmentation task for the automated diagnosis process of this disease. Then, we will present the related deep learning background. Especially, we will talk about convolutional neural network based methods. Finally, we will present the related literature in the domain of histopathological images.

2.1 Domain Background

According to the statistics in 2019, colorectal cancer is the third most common cancer type diagnosed in the US [22]. Also, more than 90 percent of all colorectal carcinoma cases are colon adenocarcinomas [23]. Colon adenocarcinoma originates from epithelial cells, which are responsible from secreting substances (e.g., hormones and mucus) and absorbing useful materials from waste products. Glands are composed of the epithelial cells surrounding around large white areas, lumens. The nuclei of the epithelial cells lie in the boundaries of the gland. The region between the glands consists of stroma tissue, which plays a connective and supportive role for the glands. For an illustrative example, see Figure 2.1.

There are some screening tests (e.g., colonoscopy and sigmoidoscopy) that facilitate the early detection of colon adenocarcinoma. However, the diagnosis of this disease and the selection of the appropriate treatment involve the histopathological examination as an essential final step. This step requires taking a colon biopsy, which is the surgical gathering of a small sample from a colon tissue. Then, after gathering, the sample is dissected into sections and stained with some chemicals to help the visual examination under a microscope. The routinely used technique is the hematoxylin-and-eosin (H&E) staining. While hematoxylin stains the nucleic acids blue providing more contrast, eosin stains cytoplasm pink providing more highlight [24] (see Figure 2.1). The final examination is done by pathologists. Since the number of cancer patients are high, the examination process is very time consuming, and the examinations done by individual pathologists are subjective, use of computer-based methods to automate this process is worthwhile [1].

Colon adenocarcinoma distorts the distribution of the epithelial cells in the glands, and thus, the morphological characteristics of the glandular structures. As a result, for detection and grading of this cancer, the quantification of the distortion level is important. To obtain this quantification with a computer-based method, extraction of the morphological characteristics of the glands is required. The first step for this is the localization of the glands by delineating

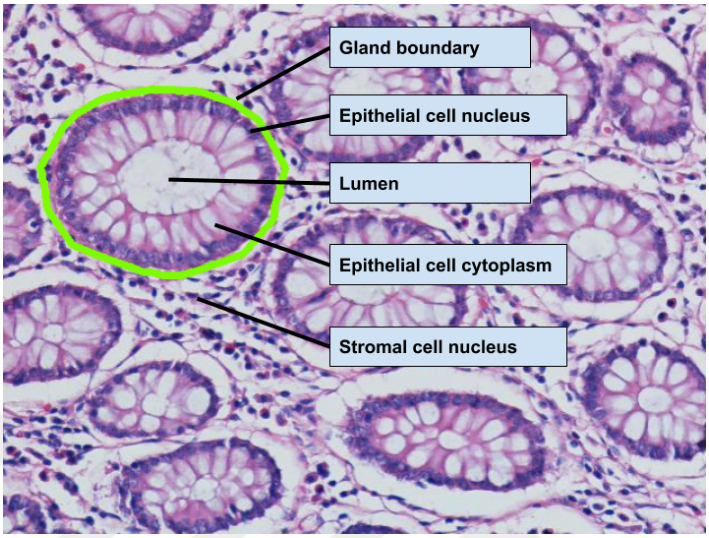


Figure 2.1: A colon tissue sample stained with the routinely used hematoxylin-and-eosin (H&E) technique.

their boundaries by a segmentation method. Since the characteristics needs to be calculated for individual glands, the segmentation method should detect not only glandular pixels, but individual gland instances as well. Therefore, gland instance segmentation is a fundamental step for automated detection and grading of colon adenocarcinoma.

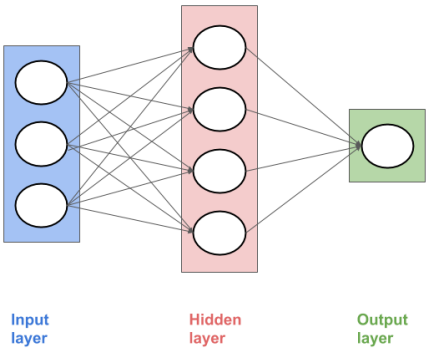
2.2 Neural Networks

An artificial neural network (ANN), one of the most well-known methods, consists of a network of fully connected elementary processors (i.e., perceptrons) operating in parallel. Based on data obtained from prior layers, each perceptron calculates a single output and the output function introduces a non-linearity in each step. The connections between the perceptrons are associated with weights. Neural networks are trained on image data in a supervised manner to learn the weights [25]. While a traditional neural network has typically one hidden layer of neurons, a deep neural network can have a much higher number of layers. Although it requires more computational power than a classical neural network, the

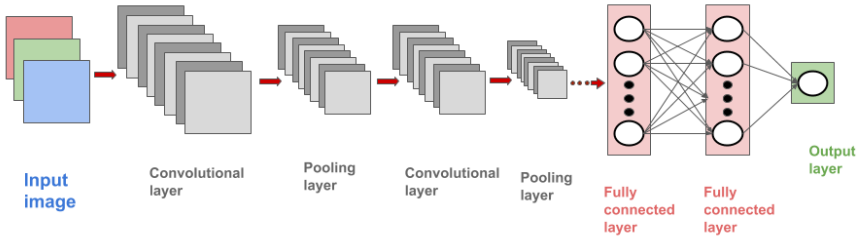
decrease in the cost of the computational power and the availability of computing methods on graphic processors (GPU) have been increasing the popularity of deep neural networks.

With the introduction of convolutional layers to the general deep neural networks architecture, convolutional neural networks (CNNs) became available. While neural networks are composed of fully connected layers only, a typical CNN architecture consists of convolutional and pooling layers followed by traditional fully connected layers at the end. The popularity of CNNs for image analysis is because of the fact that convolutional and pooling layers reduce the number of parameters significantly by providing weight sharing and enabling to use large inputs on deep networks efficiently [26]. CNNs either can be used as feature extractors taking the outputs of the fully connected layers as features, or they can directly be used as a classification model. Due to their ability to learn high-level complex features on image data [27], CNNs are demonstrated to be very successful on various image classification [28, 29, 30] and object detection [31] tasks over the last years. In the current literature, the most common deep learning techniques applied to computer vision applications are based on CNNs [19].

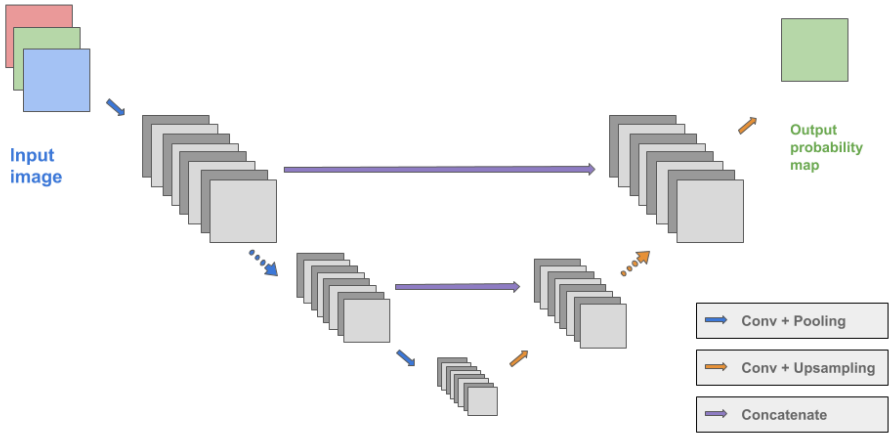
Fully convolutional networks (FCNs), proposed by Long et al. [32], are a variation of CNNs. FCN architectures are composed of only convolution and pooling layers (and possible deconvolution and upsampling layers) but they do not contain any fully connected layers at the end. The output of an FCN usually has the same size with its input. For the segmentation tasks, these dense prediction models have provided significant improvements in terms of both efficiency and accuracy. Figure 2.2 illustrates example architectures for a traditional neural network, a CNN, and an FCN.



(a)



(b)



(c)

Figure 2.2: Representative examples of different types of neural network architectures: (a) A regular 3-layer neural network, (b) a conventional CNN architecture, and (c) an FCN architecture with feature concatenations on various levels.

2.3 Gland Segmentation in Histopathological Images

Traditional techniques for gland segmentation in histopathological images comprise of techniques to extract task-specific features from images, then using those features as input to some algorithms [33]. The extracted features are also called as "hand-crafted" features and this feature extraction process can be seen as "feature engineering". For glandular and non-glandular regions, these hand-crafted features are extracted based on prior assumptions about pixel-wise intensity values [34, 35] (e.g., large white areas are lumens, darker areas are nuclei) or the spatial arrangement of some specific priors [36, 37] (e.g., glandular nuclei are at the boundaries of the lumens). In a typical framework of such techniques, initial labels are calculated by thresholding the intensity values [34, 35], using k-means clustering [36], or decomposing the image into superpixels [37]. Then, final results are obtained by using a region growing algorithm with assigned initial seeds on the thresholded map [34, 35] or by constructing spatial-arrangement graphs and using the features of these graphs to select and grow the luminal regions [36, 37]. Since their success heavily depends on the efficiency of the extracted features, these traditional techniques require paying utmost attention to the feature extraction process, trying to obtain features to represent a real-world problem as good as possible.

More recently, deep learning techniques present the advantage of learning features as higher level abstractions from the input directly and not requiring any assumption on the specific task and the dataset [38]. Thus, the feature discovery ability of deep learning has solved the drawbacks of the traditional approaches, by decreasing the required effort for task-dependent assumptions and domain specific issues on the feature extraction process.

With the advances in deep learning techniques, studies in the gland instance segmentation domain have focused on designing/employing neural network architectures suitable for the task. There have been many studies using models based

on a convolutional neural network (CNN) architecture for the gland segmentation task. Earlier studies employ the "sliding window approach", also called as "patch classification", using a conventional CNN architecture for segmentation of colon glands [39, 40, 41]. In this approach, to obtain the classification label of a pixel, an extracted patch around that pixel is fed into the network; to obtain the segmentation mask for an image, predictions for all patches are aggregated. In two studies [41, 39], three different types of classifiers have been trained and their results are compared: a support vector machine (SVM) with hand-crafted features, an SVM with features extracted from a CNN, and a CNN as the classifier. Xu et al. [41] have used a CNN model consisting of two convolutional and pooling layers, two fully connected layers, and a softmax activation function. While this relatively shallow model is trained from scratch [41], Li et al. [39] have trained deeper models with transfer learning and fine-tuning approaches using both the AlexNet [28] and GoogleNet [30] models. The findings of both show that CNN models, whether it is only to extract features or as the main classifier, are significantly better than an SVM with hand-crafted features on the gland segmentation task. From the methods that use the "sliding window approach", only the model of Kainz et al. [40] has taken the problem as segmenting individual gland instances rather than identifying gland pixels. They use two different CNNs for this purpose: one for classification of benign gland, malignant gland, and background areas, and the other for classification of gland separating areas. They use additional manual annotations for the second CNN. They combine the results and regularized them by their weighted total variation method [40].

Although the sliding window approach with conventional CNNs has outperformed the classifiers with hand-crafted features, this approach has still some drawbacks. Firstly, while larger patches provide more information, as the size of the patches increases, the number of parameters and memory requirements increase dramatically. Additionally, in this approach, extracted patches around each pixel is used to train the network and to make pixel-wise classifications. However, these patches have many overlaps resulting in redundant computations.

Fully convolutional networks (FCNs) [32] have been proposed to solve the

drawbacks of this sliding-window approach (such as reducing computational redundancy due to overlapping patches), by making the training and segmenting of larger images at once. Thus, FCN based dense prediction models have become popular architectural choices for medical image segmentation [2] with many applications including the colon gland segmentation task as well [3, 4, 12, 42, 43, 44].

From them, the U-net model, which is proposed by Ronneberger et al. [3], has become a popular choice for many segmentation tasks. This network can be divided into two symmetric paths: a contracting path with convolution and max pooling layers, and an expansive path with convolution and upsampling layers. Thus, it provides the same resolution output size with the input. It also adds skip-connections, which is also proposed by [45], between various feature channels of the contracting path and the expansive path, making the network able to use multi-level feature information without the limitations of the constant receptive field size. Since the U-net model is for the instance segmentation task, in order to make the model able to learn boundary pixels well, they use precomputed loss weight maps in the training. In these weight maps, they adjust the relative loss contributions of the pixels in the border regions higher using a function based on the distances between each pixel and the boundary of the closest gland instance [3].

For the gland instance segmentation task, while some of the models use a single FCN architecture [3], some of them use multi-task [4, 42] and multi-channel approaches [5, 12]. The multi-task/multi-channel models use boundary detection as an additional task to the main segmentation task, in order to improve the ability to segment individual instances separately. They learn these two tasks together to improve the learned segmentation features with the leverage of multi-task learning and/or to use the detected boundaries to refine the results of the segmentation maps. They combine the predicted segmentation maps and the additional boundary prediction at the end, either with a simple fusion function [4] or with an additional fusion network [12]. In [5], a detection task (bounding box information) is also added as an additional task to the architecture.

As aforementioned, training deep learning architectures for the gland instance

segmentation task is difficult without further adjustments, because of the absence of large datasets and uneven distribution of pixels of the different characteristics in background and foreground classes. If no adjustments were made, the networks would learn poor generalizations for pixels of a minority class as well as for hard-to-learn pixels. A typical case that is mostly taken into consideration by the previous models is the difficulty of being able to classify the boundary pixels accurately. It is an important challenge for the instance segmentation task, since the boundary pixels separate multiple gland instances from each other and the success of classification of the boundary pixels greatly affects the success of the entire instance segmentation task. Yet, the total weight contribution of such hard-to-learn pixels to the loss function is relatively low because of their low number of occurrences in both foreground and background classes. To solve this problem, the aforementioned FCN based models explicitly define their model's attention before training the network, by either incorporating precomputed loss weights [3] or defining additional tasks to the training process [4, 12, 5, 42]. Although both of these approaches may help handle this single problem type, namely "incorrect boundary classification", and provide better separation of touching components, there may exist other problem types associated with other hard-to-learn pixels in the images (see Figure 1.1). To make these approaches scalable to multiple types of problems, manual and external identification of each type before designing/training a network is required either by using new weight adjustments or by defining new additional tasks. Since these problems might be related with noise and artifacts, but not the nature of the images, this might be challenging. In the light of these issues, this thesis proposes a new error-driven multi-attention model, *AttentionBoost*, which adaptively learns what to attend directly on image data without making any prior assumption.

There are a limited number of studies focusing on the gland segmentation task. Thus, in the following subsection, we review other deep learning architectures related to our proposed model, even though they are not designed for the gland segmentation problem.

2.4 Other Related Network Architectures

Multi-stage FCN models have been proposed to improve the predictions of a single model [6, 7, 8, 9, 10]. These multi-stage models are combinations of multiple iterative stages, each of which learns to refine the prediction map of the previous stage by using the image and the previous map, starting with a null label map [7, 8] or a segmentation map obtained from another model [9, 10], as inputs. After some number of iterations, they use the last prediction map as the final result. The premise of these models is that learning image features together with high-level context features from the previous segmentation map will implicitly improve the results at each stage. Thus, in these models, each model in each stage learns the same task with the same objective function. Although the proposed *AttentionBoost* model is constructed as a multi-stage architecture similar to these models, as opposed to these models, the *AttentionBoost* model adaptively adjusts the objective function from one stage to another and forces the network at each stage to change its attention to learning incorrectly segmented pixels.

In the literature, different weighting strategies for the objective functions are proposed to tackle the class imbalance problem, such as weighted cross-entropy loss function [13, 14] or weighted Dice loss function [15]. In this regard, they use a constant predefined weight for all predictions of the pixels in the same class. In [13, 14], "median frequency balancing" is used to select those constant weights. In this approach, the weight of each class is the ratio of the median of all class frequencies divided by the class frequency. Thus, while the weights of the more frequent classes are smaller than 1, those of less frequent classes become higher than 1 increasing the given attention to the minority classes. In [15], to weight the Dice loss, the constant weights are selected as inverse of the volume of the classes, to decrease the effect of the region size to the Dice score. Different than these studies, the proposed *AttentionBoost* model adjusts the weights in the loss function allowing different weights for the pixel-wise predictions of the same class. The "focal loss" by Lin et al. [16] is the only proposed strategy that attempts to adjust loss weights for an FCN model during training. The strategy proposed by this study is to train a single stage model for the object detection task on

image data. During the training of this model, the loss weight for each individual object is updated at each epoch and the next epoch uses the updated weight in every pixel within the same object bounding box. By doing so, the model might increase the importance of misdetected and more difficult objects for learning in later epochs. However, because it uses a single network, the common type of incorrectly/correctly detected objects may still overwhelm the loss function, which makes it difficult to explicitly concentrate on several detection subtasks with different levels of difficulty at the same time. In addition, this study [16] uses the same loss weight for all pixels of the same object (bounding box), without any consideration of being given to their pixel-wise contributions. On the other hand, the *AttentionBoost* model is constructed with multiple stage networks, each of which can have a different attention for each pixel individually. This enables each stage to better concentrate on a different aspect of the task, depending on the difficulty of learning individual pixels.

The literature also contains studies that use the Adaboost algorithm [11] with a neural network architecture [17, 18, 19, 20, 21]. Schwenk and Bengio [17] present one of the first examples analyzing a simple neural network architecture trained with different boosting approaches (e.g., resampling the training images and weighting the cost function). The Adaboost algorithm is also used in terms of incremental learning of multiple CNNs to select a subset of samples [18] or a subset of features [21] to be used at each additional network. Gao et al. [19] employ the approach for the sentiment analysis task to boost the classification performance of CNNs. Rather than training multiple networks, a boosting like algorithm is utilized by [20] to select the sample weights during training a single CNN architecture for the pedestrian detection and action recognition tasks. However, these previous studies have been designed for the classification of an image and different than our model. They do not include a dense prediction task using an FCN. Thus, for each image, they use the same attention either by arranging different sets of training images for each learner or by arranging loss weights for each learner's training images. The *AttentionBoost* model uses the idea to adjust pixel-wise loss weights of a dense segmentation model.

Chapter 3

Methodology

The definition of the task, and hence, the objective function greatly affects the success of a network, which is trained to optimize this objective function. If the training dataset has imbalanced distributions and all data points contribute uniformly to the objective function, the network is biased to learning the most common patterns in the dataset. In this case, less common patterns need to be emphasized in the learning process. It might not, however, be simple to modify a model that contains a single network with a single objective function for many different patterns. On the other hand, it is easier to make these changes if the model allows multiple (sub)networks to be trained using different objective functions, because this facilitates the model to modulate the attention of each network with a different emphasis on the goal. Also, when multiple networks are used, each subnetwork's focus on the goal can be adjusted automatically by depending on each other.

With this motivation, the *AttentionBoost* model proposed by this thesis aims to develop a multi-stage dense segmentation network which gives specific attention to correct its mistakes automatically. For this purpose, it proposes an attention learning mechanism for this dense multi-stage prediction model, in which the attention of each stage is automatically adjusted. In this new loss adjustment mechanism, the loss contribution of each pixel prediction at each stage is

adjusted based on the level of confidence on its correct and incorrect predictions in previous stages. In order to achieve a final segmentation result, the outputs of all these stages are combined. More details about the proposed multi-stage network architecture, the attention learning method, and the inference procedure are provided in the following subsections.

3.1 Multi-Stage Network Architecture

The proposed *AttentionBoost* model is a multi-stage network that contains four segmentation subnetworks (FCNs) with a different loss attention in each stage. This multi-stage network's architecture is shown in Figure 3.1. In this multi-stage network architecture, iteratively at each stage, the nth segmentation subnetwork, takes a normalized RGB image I and a probability map $\widehat{\mathcal{Y}}_{n-1}(I)$ estimated by the previous stage as input and produces a new probability map $\widehat{\mathcal{Y}}_n(I)$ for the next stage. For all the segmentation subnetworks at each stage of the *AttentionBoost* model, we have used the same base model architecture for simplicity. In order to employ the same base model for all stages, a null map is used for $\widehat{\mathcal{Y}}_0(I)$ where $\hat{y}_0(p) = 0.5$ for all pixels. It must be noted that, since the segmentation subnetworks are trained as separate models (without weight sharing), they can be constructed as different FCN architectures as well.

The base model architecture is constructed as an FCN that has two symmetric paths (contracting path and expansive path) with feature concatenations (skip-connections) at various levels, similarly with the U-net [3] model. This architecture has the convolution layers with a 3×3 filter size and pooling/upsampling layers with a 2×2 filter size. It uses rectified linear unit (ReLU) nonlinearity on all of the convolution layers except the last one and sigmoid activation function is used in the output layer. Different than the U-net model [3], this base model uses dropout regularization [46] to reduce overfitting. Also, the number of output filters in [3] is reduced to half in each convolution for the sake of efficiency, with the ability to receive the support of GPU for the multi-stage version during training and inference. This base model architecture is shown in Figure 3.2.

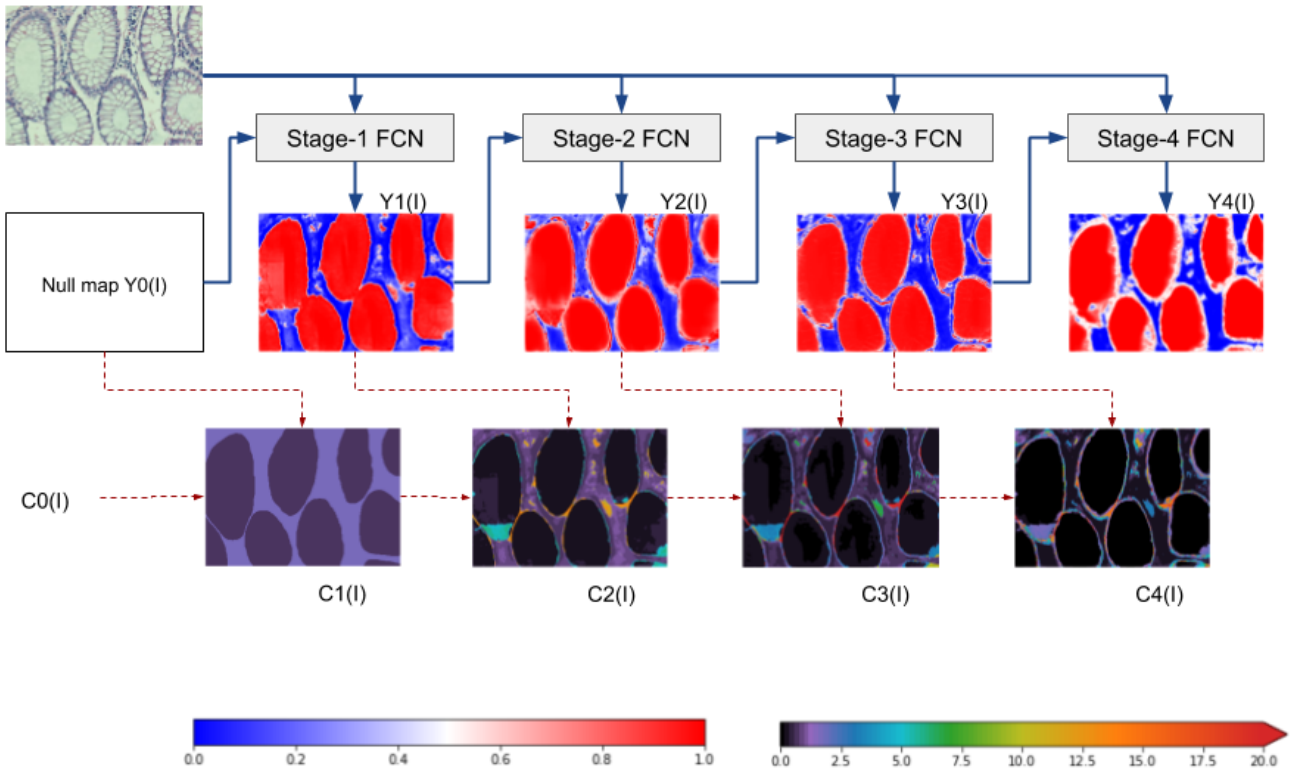


Figure 3.1: Illustration of the proposed multi-stage network architecture that consists of four segmentation subnetworks (FCNs). The n-th stage subnetwork inputs an image I and a probability map $\widehat{\mathcal{Y}}_{n-1}(I)$ estimated by the previous stage and outputs a new probability map $\widehat{\mathcal{Y}}_n(I)$. While training the multi-stage network, the loss contribution of each pixel prediction of each stage n is adjusted by the loss contribution map $\mathcal{C}_n(I)$. In order to illustrate how this multi-stage network iteratively corrects its errors for an unseen image, a test set image is used for this figure. Note that the loss contribution maps $\mathcal{C}_n(I)$ of this test image are calculated just for a demonstration purpose. The color bars given at the bottom shows the equivalent values of the colors in the illustration of the posterior maps (left) and the contribution maps (right).

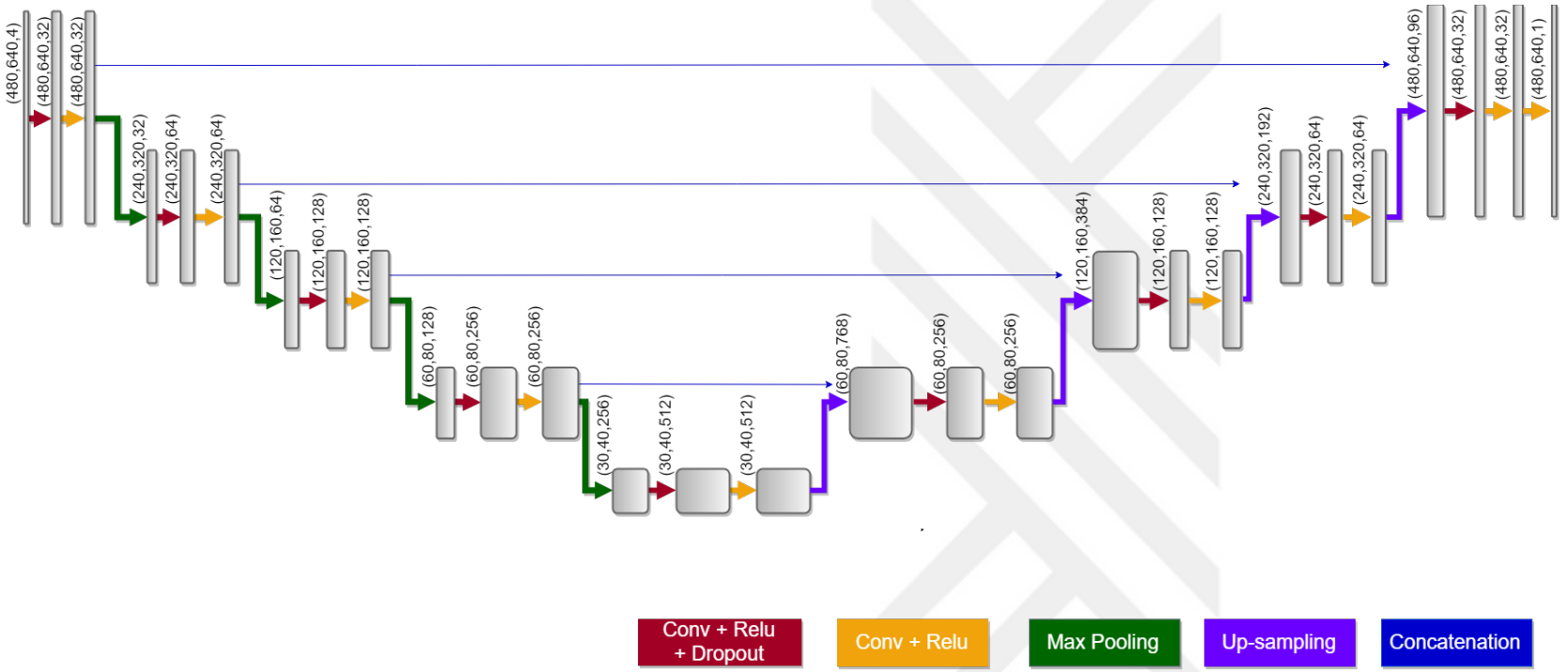


Figure 3.2: Architecture of the FCN used as the base model. This architecture consists of a contracting and an expansive path that are connected by symmetric connections, similar to [3]. Each box represents a feature map with its dimensions and number of channels being indicated in order on its right. Each arrow corresponds to an operation which is distinguishable by its color.

3.2 Multi-Stage Network Training with Attention Learning

The proposed model consists of multiple stages that use the sum of the squared errors of the predictions of image pixels multiplied by weights calculated with our proposed adjustment method. This loss function $\mathcal{L}_n$, defined for the n-th stage network, is given as follows:

$$\mathcal{L}_n = \sum_{I \in \mathcal{D}} \sum_{p \in I} C_n(p) \cdot \Big(y(p) - \hat{y}_n(p) \Big)^2 \tag{3.1}$$

The notation used in this equation is:

- I: an image in the training set $\mathcal{D} = \{I, \mathcal{Y}(\mathcal{I})\}$ where $\mathcal{Y}(I) = \{y(p)\}_{p \in I}$
- p: a pixel in the training image I
- $y(p)$: the ground truth for pixel p. Here, the ground truth $y(p)$ is defined as 1 if the pixel belongs to a gland region and as 0 otherwise.
- $\hat{y}_n(p)$: the probability estimated for pixel p by the n-th stage network
- $C_n(p)$: the contribution of this pixel prediction to the loss function $\mathcal{L}_n$

The attention learning mechanism of the *AttentionBoost* model provides for a simultaneous learning of these contributions $C_n(p)$, for each pixel p and for each stage n, with the learning of the network weights by backpropagation. Specifically, this mechanism reduces the loss contributions of pixels if they are correctly estimated by the previous stage and increases these contributions if the pixels are incorrectly estimated by the previous stage, in the context of adaptive boosting.

To do so, we define the $\beta_n(p)$ coefficient that will determine how much effect the current loss contributions $C_n(p)$ will have on the loss contributions of the next stage $C_{n+1}(p)$. Provided that the initial loss contributions $C_0(p)$ are predefined

as 1 or depending on some prior information such as the class distributions, we can compute the coefficient value for the next stage as follows:

$$C_{n+1}(p) = \beta_n(p) \cdot C_n(p) \tag{3.2}$$

$$\beta_n(p) = \begin{cases} 1 - |\hat{y}_n(p) - 0.5| & \text{if } \hat{y}_n(p) \text{ is correct} \\ 1 + |\hat{y}_n(p) - 0.5| & \text{if } \hat{y}_n(p) \text{ is incorrect} \end{cases} \tag{3.3}$$

The $C_n(p)$ values computed by the given equations are used to weigh the mean squared error loss of the corresponding stage n during the model training. Here, it can be seen that the $|\hat{y}_n(p) - 0.5|$ term in Equation 3.3 corresponds to the confidence level of the n-th stage network on its estimation for pixel p and it holds that $0 \leq |\hat{y}_n(p) - 0.5| \leq 0.5$. Thus, the resulting $\beta_n(p)$ will be the maximum 1.5 if the estimation is incorrect but very confident, and the loss contribution $C_{n+1}(p)$ will become larger forcing the next stage network to increase its attention to learning pixel p. Contrastly, $\beta_n(p)$ will be minimum 0.5 if the estimation is correct and very confident. This makes the loss contribution $C_{n+1}(p)$ smaller forcing the next stage network to decrease its attention to learning pixel p. Thus, with these attributes, the $\beta_n(p)$ coefficients are used to adjust the attention of the next stage $n+1$ for the pixel p.

After calculating the loss contributions $C_{n+1}(p)$ using Equation 3.2, these contributions are normalized such that:

$$\begin{aligned} \sum_{p \in G_n} C_{n+1}(p) &= \frac{W \times H}{2} \\ \sum_{q \notin G_n} C_{n+1}(q) &= \frac{W \times H}{2} \end{aligned} \tag{3.4}$$

where $G_n = \{\ p \in I |\ \hat{y}_n(p) \text{ is correct}\}$. This normalization ensures that the next stage networks will not totally abandon their attention to the accurate segmentation of pixels, since the sum of the coefficients $C_{n+1}(p)$ for the correctly estimated

pixels p in image I, and the sum of the coefficients for all incorrectly estimated pixels q in image I are equal. This is essential because at the end the final segmentation is achieved by adding the output maps of all stages (Section 3.3). In these equations, $W \times H$ represent the dimensions (number of pixels) of the input image I, and it is to scale the learning rate and to avoid having very small gradients.

During the end-to-end training of the proposed *AttentionBoost* model, the normalized RGB images I in the training set $\mathcal{D}$ and the ground truth segmentation maps $\mathcal{Y}(I) = \{y(p)\}_{p \in I}$ are given to the network. In the forward pass, the loss contributions $\mathcal{C}_n(I) = \{c_n(p)\}_{p \in I}$ for each training image I at each stage n is calculated and the loss functions $\mathcal{L}_n$ are updated accordingly. In the backward pass, the network weights are updated derivating these updated loss functions.

3.3 Gland Segmentation

This step combines the segmentation (posterior) maps generated by all stages of the proposed *AttentionBoost* model. Since the *AttentionBoost* model is a multi-stage and an error-driven model, its stages are expected to produce complementary maps, especially for hard-to-learn pixels. As different types of hard-to-learn pixels have different characteristics, it is hard for a single stage network to produce correct predictions for all of these pixels. On the other hand, by forcing each stage to learn the mistakes of the previous stages, one stage is expected to compensate the errors of the previous stages. With this strategy, a more balanced learning of the task and more robust predictions are achieved. To combine these maps, we utilize a straightforward approach, although it may be considered to design and use more advanced methods to process them. The approach used in this thesis calculates the average of all probability maps and then applies a region growing algorithm on this average map. In Figure 3.3, gland segmentation for a test set image I is given along with the posterior maps obtained from all stages and the ground truth segmentation of this image. In this figure, posteriors between 0.5 and 1 (these are the posteriors of pixels belonging to the gland class)

are shown in red, and those between 0 and 0.5 are shown in blue. The darker the tone of the red(blue) color is the more confident the corresponding network is on its prediction. As can be seen in this figure, each stage corrects different errors of the previous stages (while it can also yield new errors), and as a result, the average posterior map becomes more balanced and accurate than each of the stages.

To obtain a final segmentation map for image I, the image is first given to the trained network as input and the output probability maps $\hat{\mathcal{Y}}_n(I) = \{\hat{y}_n(p)\}_{p \in I}$ from each stage n are aggregated by taking the average. Based on pixel prediction $\hat{y}_{avg}(p)$ in this average probability map $\widehat{\mathcal{Y}}_{avg}(I) = \{\hat{y}_{avg}(p)\}_{p \in I}$ a label $l(p)$ is given to each pixel p as follows:

$$l(p) = \begin{cases} \text{foreground} & \text{if } \hat{y}_{avg}(p) \geq 0.5 + \alpha \\ \text{background} & \text{if } \hat{y}_{avg}(p) \leq 0.5 - \alpha \\ \text{uncertain} & \text{otherwise} \end{cases} \tag{3.5}$$

Here the given label $l(p)$ indicates whether or not the pixel p certainly belongs to a foreground or a background region, depending on a confidence parameter α. Thus, the "certain" foreground, "certain" background regions and "uncertain" pixels are identified. Then, the connected components in the "certain" foreground and "certain" background regions are found seperately, and components smaller than an area threshold A_{thr} are eliminated. After adding the pixels of the eliminated regions to the "uncertain" pixels, the segmentation maps are obtained by growing "certain" regions onto "uncertain" pixels with respect to the average probabilities. As the final step, a majority filter with a size of f_{size} is applied to these segmentation maps to smooth the boundaries.

.

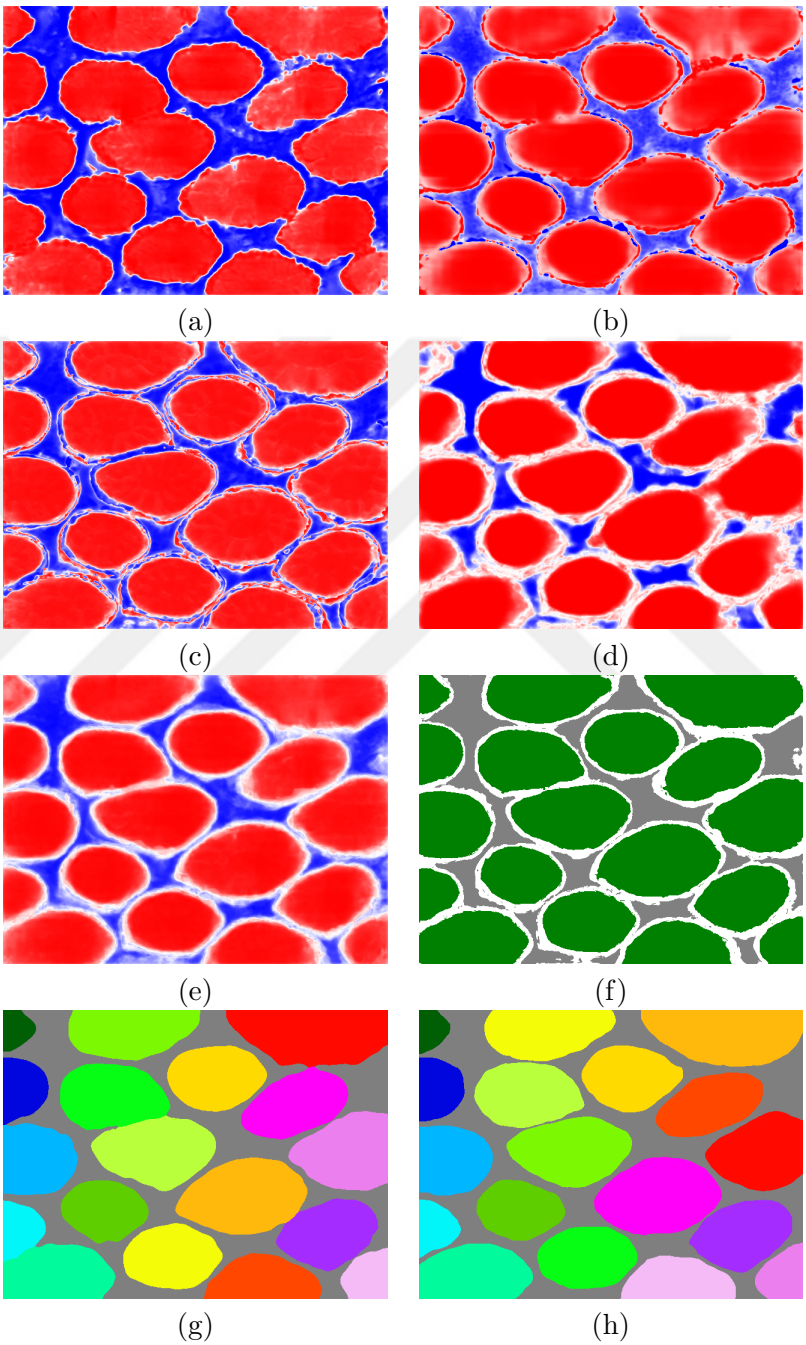


Figure 3.3: Gland segmentation for a test set image I. (a)-(d) Posterior maps $\widehat{\mathcal{Y}}_i(I)$ for first, second, third, and fourth stage respectively. (e) Average probability map $\widehat{\mathcal{Y}}_{avg}(I) = \{\hat{y}_{avg}(p)\}_{p \in I}$. (f) Label map $L(I) = \{l(p)\}_{p \in I}$ where the "certain" foreground (green), "certain" background (gray), and "uncertain"(white) pixels are identified. (g) Final segmentation result. (h) Ground truth segmentation.

Chapter 4

Experiments

4.1 Dataset

We test our model on a dataset containing 200 microscopic images of colon biopsy samples from the Hacettepe University School of Medicine Pathology Department Archives. The samples are tissue sections stained with the routinely used hematoxylin-and-eosin staining. They contain both normal and cancerous (colon adenocarcinomatous) glands. The images of the samples are taken by a Nikon Coolscope Digital microscope with a 20× objective lens. The the image resolution is 480×640 pixels.

The dataset is split into training, validation, and test sets. The number of images and the number of glands in each set are given in Table 4.1. The backpropagation algorithm uses the training images to learn the weights of the proposed multi-stage network and the validation images are used to stop the backpropagation algorithm. The training and validation images are used to choose the confidence parameter α, the area threshold A_{thr}, and the majority filter size f_{size} by the gland segmentation step. The parameter selection is explained in Section 4.4. The test images are not used for network training or parameter selection, but only for evaluation purposes.

Table 4.1: Number of images and number of glands in the training, validation, and test sets.

	Number of images			Number of glands		
	Training	Validation	Test	Training	Validation	Test
Normal	40	10	50	570	174	621
Cancerous	40	10	50	321	49	367
Total	80	20	100	891	223	988

4.2 Implementation Details

The proposed multi-stage network and attention learning mechanism is implemented in Python using the Keras deep learning framework [47]. The network is trained on a GPU (GeForce GTX 1080 Ti) using randomly initialized network weights from scratch and with an early stopping strategy depending on the loss of the validation images. The batch size is 1 and the drop-out rate of all drop-out layers is equal to 0.2. The AdaDelta optimizer [48] is employed for the gradient descent to adaptively adjust the learning rate.

4.3 Evaluation

Segmentation results are quantitatively evaluated using object-level F-score, the object-level Dice index, and the object-level Hausdorff distance. Those three criteria are explained in the following subsections:

4.3.1 Object-Level F-score

The object-level F-score is used for an assessment of the correctly detected percentage of gland objects. The object-level F-score is defined as:

$$\begin{aligned} \textit{F-score} &= \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} \\ \textit{precision} &= |TP|/(|TP| + |FP|) \\ \textit{recall} &= |TP|/(|TP| + |FN|) \end{aligned} \tag{4.1}$$

Considering segmented gland objects and ground truth objects, true positive (TP), false positive (FP), and true negative (TN) objects are identified as follows:

- True positive (TP): A segmented gland object for which intersection with any ground truth object is greater than 50 percent of this ground truth object.
- False positive (FP): A segmented gland object that is not a true positive.
- False negative (FN): A ground truth object which does not match with any true positive segmented gland object.

4.3.2 Object-Level Dice Index

The object-level Dice index is to measure how precisely the pixels in the segmented gland objects overlap with the pixels in their matched (maximally overlapping) ground truth objects. The Dice index between two objects A and B is defined as follows:

$$DI(A, B) = \frac{2 \cdot |A \cap B|}{(|A| + |B|)} \tag{4.2}$$

To calculate the object-level Dice index for a given segmentation $S = \{s_i\}$ with respect to the ground truth $G = \{g_j\}$, this Dice index is calculated for every set of matching objects; one from the set of segmented gland objects $S = \{s_i\}$

and the other from the set of ground truth objects $G = \{g_j\}$. To do so, matching objects (a segmented gland object s_i and matching ground truth object $\gamma(s_i)$, and similarly a ground truth object g_j and matching segmented gland object $\sigma(g_j)$) should be identified. The overlapping regions should be maximum between the matched objects. If an object is not matched, the contribution of this object to the object-level Dice index is zero. Then, the object-level Dice index is a weighted sum of all Dice indices defined for object pairs. It is defined as follows:

$$Dice(S,G) = \frac{1}{2}\left(\begin{array}{c} \sum\limits_{s_i \in S} \omega(s_i) \cdot DI(s_i, \gamma(s_i)) \\ + \\ \sum\limits_{g_j \in G} \omega(g_j) \cdot DI(g_j, \sigma(g_j)) \end{array}\right) \tag{4.3}$$

where $\omega(s_i) = |s_i| \ / \sum\limits_{s_m \in S} s_m$ and $\omega(g_j) = |g_j| \ / \sum\limits_{g_m \in G} g_m$ are the weights that determine the contribution of each Dice index to the final score.

4.3.3 Object-Level Hausdorff Distance

The Object-level Hausdorff distance evaluates the shape similarity between the segmented gland objects and their matching ground truth objects. The Hausdorff distance between two objects A and B is calculated as follows:

$$HD(A,B) = \max\{\sup_{p_A \in A} \inf_{p_B \in B} ||p_A - p_B||, \sup_{p_B \in B} \inf_{p_A \in A} ||p_A - p_B||\} \tag{4.4}$$

where $\sup\limits_{p_A \in A} \inf\limits_{p_B \in B} ||p_A - p_B||$ is the maximum of the minimum distances calculated from every pixel p_A of the object A to any pixel p_B of the object B.

While calculating the Hausdorff distance, for each s_i and g_j the matching objects $\gamma(s_i)$ and $\sigma(g_j)$ are identified similarly with the object-level Dice index. The only difference is that, if an object does not have any overlapping counterpart, this object is matched with the object that has the minimum Hausdorff distance from it. Then, the object-level Hausdorff distance is the weighted sum of all the Hausdorff distances for matching object pairs. It is defined as follows:

$$Hausdorff(S,G) = \frac{1}{2}\left(\begin{array}{c}\sum\limits_{s_i \in S} \omega(s_i) \cdot HD(s_i, \gamma(s_i)) \\ + \\ \sum\limits_{g_j \in G} \omega(g_j) \cdot HD(g_j, \sigma(g_j))\end{array}\right) \tag{4.5}$$

4.4 Parameter Selection

The proposed *AttentionBoost* model involves three hyper-parameters. The best combination of these parameters is selected by the grid search method according to the object-level Dice index of each combination calculated on the training and validation images. The same procedure is followed for the parameters of the comparison methods as well. Additionally, a discussion about the effects of these parameters to the model's performance is given in Section 5.2.

The hyper-parameters are described below:

- The confidence parameter α: This parameter is introduced to acquire a level of confidence on the aggregated posterior maps of different stages. Since these stages have different attentions on the images, there is an uncertainty on the aggregated posterior maps of different stages depending on how diverse the predictions of these stages for a given image are. In the grid search, we have used the values $\alpha = \{0.05, 0.10, 0.15, 0.20, 0.25\}$ for this parameter and $\alpha = 0.15$ is selected.

- The area threshold A_{thr}: Before growing the certain connected components onto uncertain pixels, the components with an area smaller than A_{thr} are eliminated and added to the uncertain pixels. This is for eliminating noisy regions. The values used for the grid search are $A_{thr} = \{250, 500, 750, 1000\}$ and $A_{thr} = 250$ is selected.

- The majority filter size f_{size}: To smooth the boundaries of the segmentation results, as a final step, the majority filter with a size of f_{size} is applied. The

set of $f_{size} = \{5, 9, 15, 19\}$ is considered by grid search and $f_{size} = 15$ is selected.

4.5 Comparisons

To be able to qualitatively and quantitatively compare the performance of the proposed *AttentionBoost* model, we have used different comparison methods based on single-stage [3, 4] and multi-stage approaches [7] in the literature. All of these approaches are for dense prediction tasks and implemented with FCN architectures. All of the parameters used in the postprocessing procedure of these models are selected by grid search on the training and validation images (see Section 4.4). For a fair comparison, we have used the same FCN architecture (the base model, Figure 3.2) for all comparison methods. These methods are explained further in the following subsections.

4.5.1 Comparison with Single Stage Approaches

For the single-stage approaches in the literature, we have used two different methods derived from [3, 4]. These methods, which we call *BoundaryAttentionWithLossAdjustment* and *BoundaryAttentionWithMultiTask*, are single-stage models and designed to give particular attention on gland borders. In contrast with our proposed model, these comparison methods define their attention as borders before the model training, and configure their system accordingly. These comparison methods are used to investigate the advantages of our proposed attention learning strategy.

4.5.1.1 *BoundaryAttentionWithLossAdjustment*

This method gives specific attention to pixels close to the gland boundaries by increasing the contribution they made to the total loss function. Before the model

training, it uses a function to adjust weights for all the pixels in an image. As explained in the U-Net model [3], this function gives higher weights to the pixels depending on their closeness to the boundary of the gland instances. In our experiments, the trained network tend to undersegment gland components. Since some of the components are linked to one another via narrow bridges in the predictions, to improve the results of this comparison method we have postprocessed its results as follows: First, the predicted connected components are eroded by a disk structuring element. Then, eroded components smaller than an area threshold are eliminated and the remaining components are dilated by using the same structuring element.

4.5.1.2 *BoundaryAttentionWithMultiTask*

This method is constructed as a multi-task architecture based on the DCAN model proposed by [4]. It gives specific attention to learning boundary pixels by using boundary prediction as an additional task to the main task of gland segmentation. The architecture of this network contains two expansive (decoder) paths using shared features of one contracting (encoder) path. After the training, the output boundary and segmentation posterior maps generated for an image are combined together and postprocessed. For this, the thresholded posterior maps are fused by subtracting the boundary map from the segmentation map. Then, on the subtracted map, connected components larger than an area threshold are found and dilated with a disk structuring element.

4.5.2 Comparison with Multi Stage Approaches

We have used a method, *MultiStageWithoutAdaptiveBoosting*, based on multi-stage approaches in the literature [7].

4.5.2.1 *MultiStageWithoutAdaptiveBoosting*

This method consists of a multi-stage model that has the same architecture with the proposed *AttentionBoost* model and it is iteratively trained. This method also generates a segmentation map at each stage using an input image and a segmentation map from the previous stage. Differently from our model, in this approach, all of the stages are trained with the same objective (loss) function without any adjustments. For this reason, this comparison method is used to understand the effect of using adaptive boosting in a dense prediction model. After the training of this comparison model, for the images in the test set, the posterior segmentation maps generated by its last stage are taken and postprocessed. The same postprocessing procedure with the *BoundaryAttentionWithLossAdjustment* method is used.

Chapter 5

Results and Discussion

In this chapter, we will present the comparison results of the proposed *AttentionBoost* model along with a discussion on the effects of the parameter selection to the model's performance.

5.1 Comparisons

To understand the performance of the proposed approach for the gland instance segmentation task, its results are analyzed both quantitatively and qualitatively, and compared with those of the other approaches. The quantitative results of *AttentionBoost* and the comparison methods are presented in the Table 5.1. For a more informative comparison, these results are obtained on all test images and on the test images containing normal and cancerous glands, separately. In turn, the resulting higher scores of the *AttentionBoost* model for the object-level F-score and object-level Dice index metrics show that the proposed approach is more successful at detection and segmentation of individual gland instances. Moreover, the lower scores of the Hausdorff distances indicate that the gland shapes in the predictions of our approach are more accurate than its counterparts. The visual results on exemplary test set images are also given in Figures 5.1 and 5.2, for

Table 5.1: Quantitative results of the proposed *AttentionBoost* model and the comparison methods obtained on the test set images.

	Normal glands		
	F-score	Dice	Hausdorff
AttentionBoost	95.39	94.58	25.89
BoundaryAttentionWithLossAdjustment	89.39	86.36	71.16
BoundaryAttentionWithMultiTask	95.59	92.48	33.51
MultiStageWithoutAdaptiveBoosting	88.50	84.04	86.08

	Cancerous glands		
	F-score	Dice	Hausdorff
AttentionBoost	91.76	92.50	42.74
BoundaryAttentionWithLossAdjustment	87.57	90.66	55.09
BoundaryAttentionWithMultiTask	84.14	89.84	46.05
MultiStageWithoutAdaptiveBoosting	90.60	91.66	50.37

	All glands		
	F-score	Dice	Hausdorff
AttentionBoost	94.03	93.56	34.12
BoundaryAttentionWithLossAdjustment	88.69	88.46	63.29
BoundaryAttentionWithMultiTask	91.13	91.20	39.61
MultiStageWithoutAdaptiveBoosting	89.31	87.77	68.62

normal and cancerous glands, respectively.

The claim of the proposed *AttentionBoost* model, differently than the comparison methods, is being able to improve the results for different types of hard-to-learn mistakes by learning what to attend in images automatically. For this reason, we also have examined the test results under three types of predetermined mistake types as well. To analyze the contribution of the proposed approach in terms of different hard-to-learn mistakes, we have used the following predetermined types of mistakes:

- *Undersegmented ground truth objects:* A ground truth object $g \in G$ is considered as undersegmented if a segmented gland object $s \in S$ intersects with at least 50 percent of g but also intersects with at least 50 percent of another ground truth object $g' \in G$.

Table 5.2: Number of the types of mistakes that the proposed *AttentionBoost* model and the comparison methods make on the test set images.

	Under-segmented	False segmented	Small over-segmented	False negative
AttentionBoost	60	15	27	42
BoundaryAttentionWithLossAdjustment	222	46	15	20
BoundaryAttentionWithMultiTask	80	55	50	30
MultiStageWithoutAdaptiveBoosting	215	16	16	31

- *False positives:* A segmented gland object $s \in S$ is considered as false positive if it does not intersect with at least 50 percent of any ground truth object $g \in G$. Also, we call this false positive s as:
  - *False segmented object* : if any $g' \in G$ does not intersect with at least 50 percent of s.
  - *Small oversegmented object* : if a ground truth object $g' \in G$ intersects with at least 50 percent of s.
- *False negatives:* A ground truth object $g \in G$ is considered as false negative (missing object) if at least its 50 percent does not intersect with any segmented gland object $s \in S$.

The first mistake type (*undersegmented ground truth objects*), represents the case where the labels of pixels near the gland borders cannot be predicted properly. As stated before, separation of the gland boundaries is an important subtask to succesfully segment gland instances. Because of this, this error type is widely considered by the methods in the literature. They have attempted to solve this error by changing the models' attention with predefined higher loss weights for the boundary pixels [3] or by constructing multi-task architectures specifically designed to discriminate boundaries [4, 12].

Other mistake types that greatly affect the gland instance segmentation performance are the number of false positive and false negative instances in the

predictions. In our observations, typically, false positive errors are occurred as two cases. As the first case, the models tend to segment non-gland regions around white artifacts as gland objects. These artifacts arise from the tissue preparation process (fixation and sectioning). The first row of Figure 5.1(d) contains an example of this case. The second common case where false positive errors occur is the oversegmentation of small objects in a gland, usually close to its boundary. Two examples of this kind of small oversegmented objects can be seen in the third row of Figure 5.1(c).

Table 5.2 provides the number of occurrences of each mistake type in the test set segmentation results of the methods. Figures 5.1 and 5.2 illustrate some visual examples of the results. For this gland segmentation problem, the proposed *AttentionBoost* model shows the best performance for the most common mistake types, undersegmentations and false segmented objects. While other methods have reduced either undersegmentations, which arise because border pixels are incorrectly classified, or false segmented objects, caused by the low differentiating ability between the gland and non-gland pixels, our proposed model improves the results for both at the same time. Although there is slightly more ground-truth object is missing in *AttentionBoost*, we observe that these missing ground-truth objects have small regions close to the image borders. An example of this can be seen in the last row of Figure 5.2(b) at the top-right corner of the image. The improvement on both of the mistake types at the same time is attributed to the ability of the proposed multi-stage attention learning approach to adjust different attentions at each stage depending on the errors of the previous stages. As aforementioned (in Figure 3.3), this strategy allows to produce complementing posterior maps at each stage to compensate the errors that previous stages made on hard-to-learn pixels. Figures 5.3 and 5.4 give an additional example test set image for normal and cancerous glands, respectively.

In comparison with the others we have the following observations:

- *MultiStageWithoutAdaptiveBoosting* successfully eliminates false positives, yet it produces a significantly higher number of undersegmentations. This is because its predictions at the boundary pixels is not sufficiently refined

throughout its stages. Since this model is also a multi-stage model but it is trained to optimize the same loss function at every stage, these results indicate the superiority of adjusting the attention of each stage via adaptive boosting automatically.

- Although the results of *BoundaryAttentionWithMultiTask* for undersegmentations is relatively better, this method produces more false positives, as can be seen in Figures 5.1(e) and 5.2(e). This method's efficiency at undersegmentations is due to its predefined attention using an additional border detection task in its architecture, yet, it costs a higher number of false positives in return. This shows the effectiveness of learning several attentions directly on image data, as proposed by the *AttentionBoost* model, rather than defining a specific type of attention externally in advance.

- *BoundaryAttentionWithLossAdjustment* is less effective in decreasing both undersegmented ground truth objects and false segmented glands. It appears that this model is likely to detect more glands, also leading to less missing ground truth objects.

5.2 Parameter Analysis

In the final gland segmentation step, the proposed *AttentionBoost* model uses three external parameters: confidence parameter α, area threshold A_{thr}, and majority filter size f_{size}. To evaluate the effects of each parameter on the performance of the model, the selected values of the other two parameters are fixed, then the F-score, object-level Dice index and object-level Hausdorff distance measures are calculated on the test set. These analyses are shown in Figures 5.5, 5.6, and 5.7 for these parameters, respectively.

5.2.1 Confidence Parameter α

In the gland segmentation step, for an image I, the posterior maps of all stages are aggregated, and the resulting average probability map $\widehat{\mathcal{Y}}_{avg}(I) = \{\hat{y}_{avg}(p)\}_{p \in I}$ is used to locate certain gland and background pixels. Then, the gland and background objects on these certain pixels are grown onto uncertain pixels. To select these certain pixels, the confidence parameter α is used to decide the certainty level as given in Equation 3.5.

If this parameter is selected too large, only very certain pixels, close to 0 or 1, will be selected from the average map $\widehat{\mathcal{Y}}_{avg}(I)$. This kind of average posteriors for a pixel p can be obtained only when the majority of the predictions for pixel p agrees on the same class very confidently. However, the proposed *AttentionBoost* model forces its stages to reverse the incorrect predictions on hard-to-learn pixels, and thus, produces an average map having many posteriors closer to 0.5. As a result, using larger α values leads to less certain pixels to be selected and reduces the number of gland objects to be grown. In this case, the performance of the model is very poor with lower F-scores, lower Dice indices, and higher Hausdorff distances.

Accordingly, if this parameter is chosen too small, it also leads to poor performance. Since, in this case, nearly all pixels will be labeled as certain, the correcting effect of the stages, like separating gland boundaries, will not be utilized efficiently. Boundaries of close glands typically contain pixels whose $\hat{y}_{avg}(p)$ is around 0.5 and if too low α values are used, more undersegmented gland objects will appear in the results. These analyses are presented in Figure 5.5.

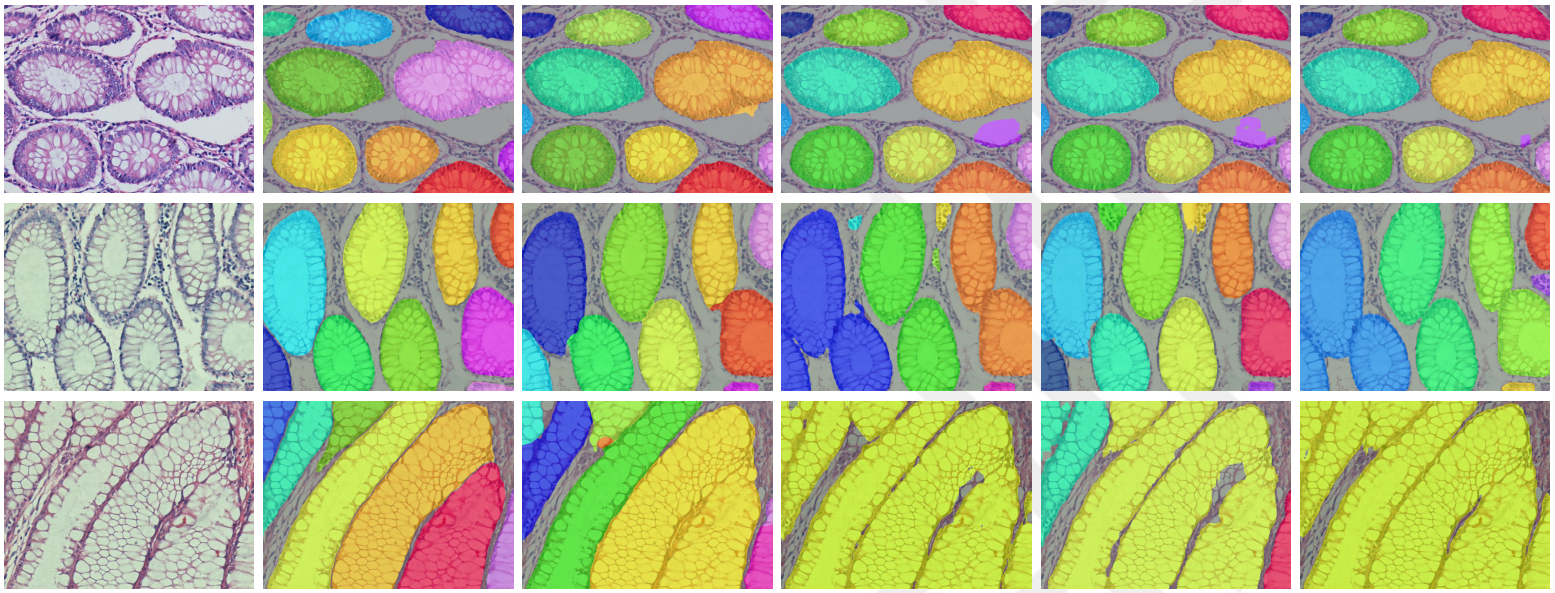
5.2.2 The Area Threshold A_{thr}

In the gland segmentation step, after certain gland and background pixels (regions) are located, those smaller than the area threshold A_{thr} are eliminated. If this A_{thr} is selected too small, the number of false positives will increase since

noisy gland objects cannot be eliminated. On the other hand, if too large values are selected for A_{thr}, the number of false negatives will increase because of the elimination of small true gland objects. Both cases, with an increased number of false positives or false negatives, will result in lower F-scores. Note that, since the object-level Dice index and object-level Hausdorff distance are weighted averages calculated depending on the areas of individual gland objects (see Equations 4.3 and 4.5), the elimination of small-sized glands does not change these measures too much. These analyses are presented in Figure 5.6.

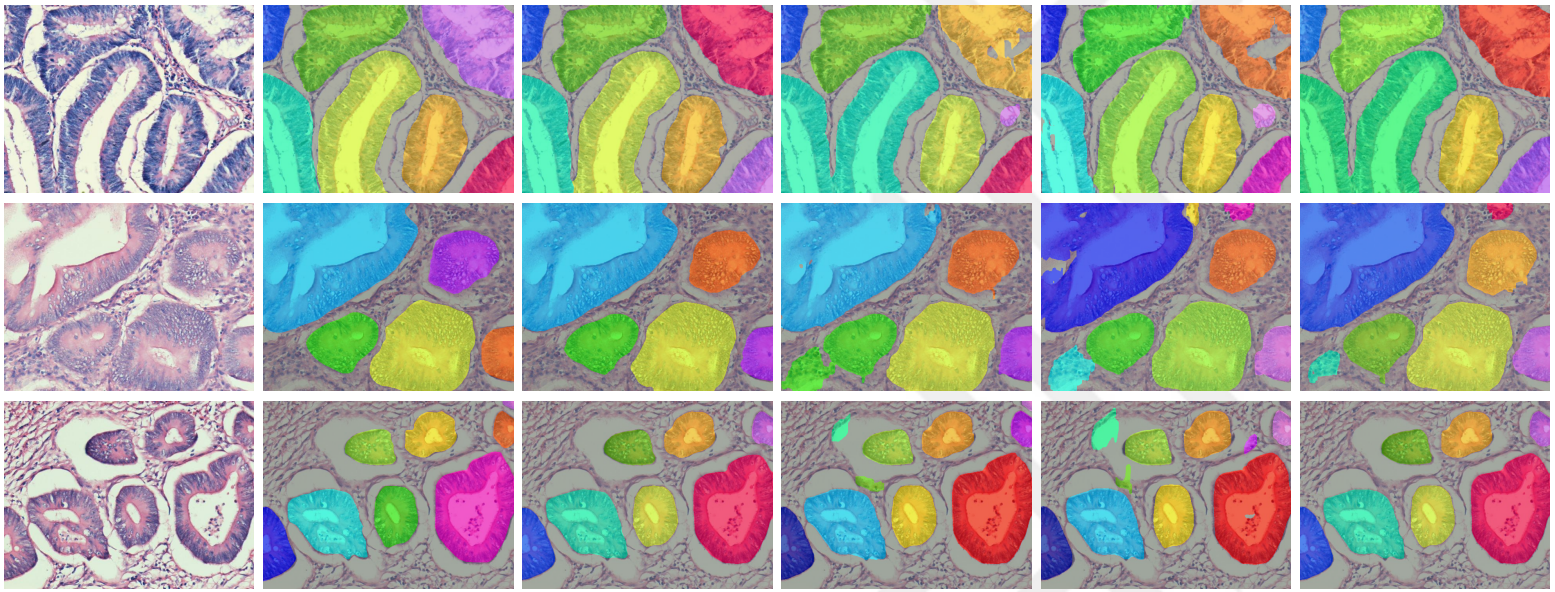
5.2.3 The Filter Size f_{size}

As the final step of the gland segmentation, the majority filter with the filter size f_{size} is applied on the grown regions. This step is only to improve the visual results by smoothing gland boundaries, it does not change the number of detected glands. Hence, its effects on the performance measures are very slight. This analysis is presented in Figure 5.7.



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

Figure 5.1: (a) Example test set images containing normal glands. (b) Ground truths. (c) Results of the proposed *AttentionBoost* model. (d) Results of the *BoundaryAttentionWithLossAdjustment* method. (e) Results of the *BoundaryAttentionWithMultiTask* method. (f) Results of the *MultiStageWithoutAdaptiveBoosting* method.



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

Figure 5.2: (a) Example test set images containing cancerous glands. (b) Ground truths. (c) Results of the proposed *AttentionBoost* model. (d) Results of the *BoundaryAttentionWithLossAdjustment* method. (e) Results of the *BoundaryAttentionWithMultiTask* method. (f) Results of the *MultiStageWithoutAdaptiveBoosting* method.

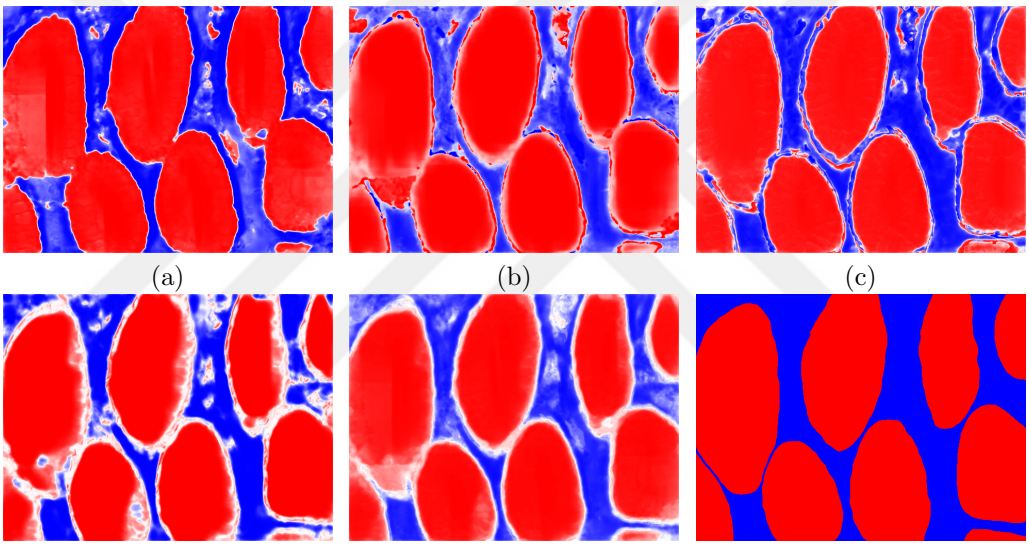


Figure 5.3: Segmentation (posterior) maps illustrated for a test set image containing normal glands. (a)-(d) Posterior map $\widehat{\mathcal{Y}}_n(I)$ generated by the first, second, third, and fourth stage, respectively. (e) Average posterior map $\widehat{\mathcal{Y}}_{avg}(I)$ obtained by aggregating the posterior maps of all stages. (f) Posterior map $\mathcal{Y}(I)$ for the ground truth segmentation. In these maps, posteriors between 1 and 0.5 (these are the posteriors of pixels belonging to the gland class) are shown in red, and posteriors between 0 and 0.5 are shown in blue. The darker the color is, the more confident the prediction is. In these maps, posteriors close to 0.5 seem whitish.

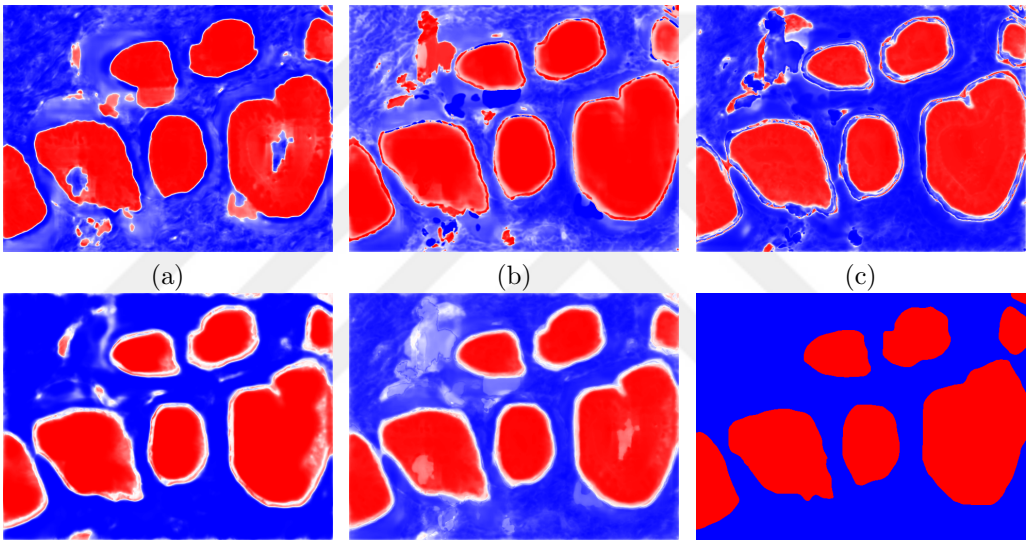
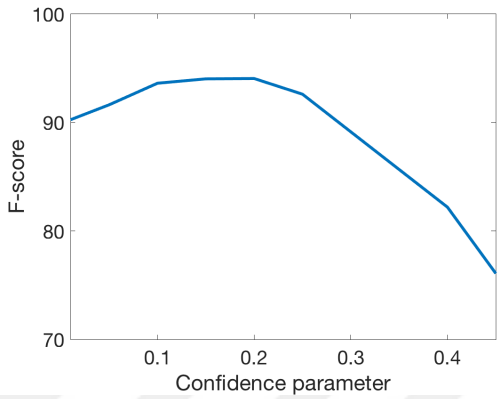
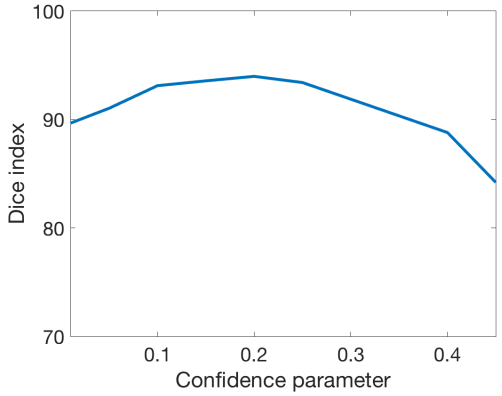


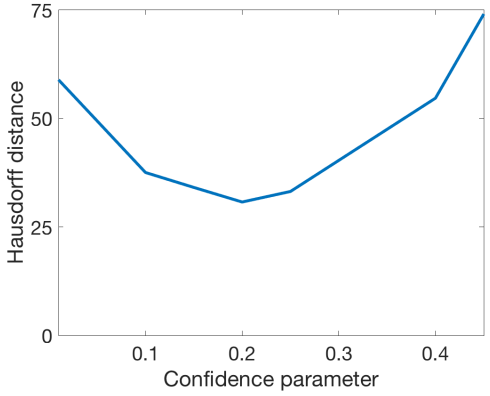
Figure 5.4: Segmentation (posterior) maps illustrated for a test set image containing cancerous glands. (a)-(d) Posterior map $\widehat{\mathcal{Y}}_n(I)$ generated by the first, second, third, and fourth stage, respectively. (e) Average posterior map $\widehat{\mathcal{Y}}_{avg}(I)$ obtained by aggregating the posterior maps of all stages. (f) Posterior map $\mathcal{Y}(I)$ for the ground truth segmentation. In these maps, posteriors between 1 and 0.5 (these are the posteriors of pixels belonging to the gland class) are shown in red, and posteriors between 0 and 0.5 are shown in blue. The darker the color is, the more confident the prediction is. In these maps, posteriors close to 0.5 seem whitish.



(a)

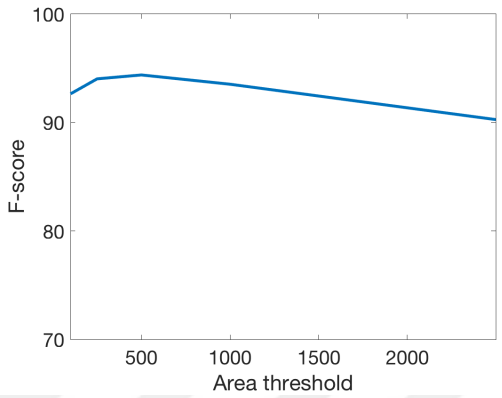


(b)

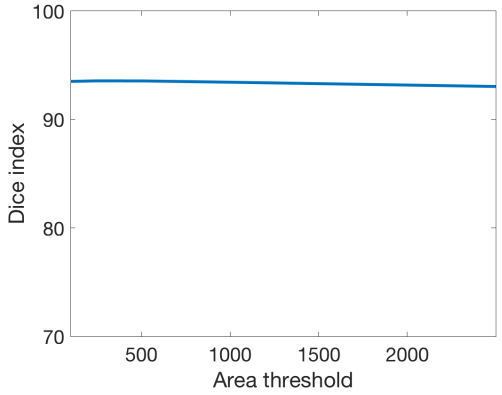


(c)

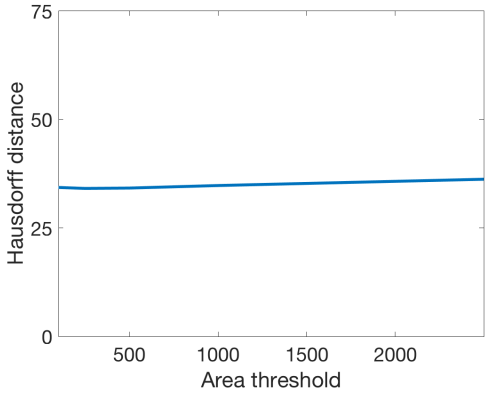
Figure 5.5: Test set F-scores, object-level Dice indices, and object-level Hausdorff distances as a function of the confidence parameter α.



(a)

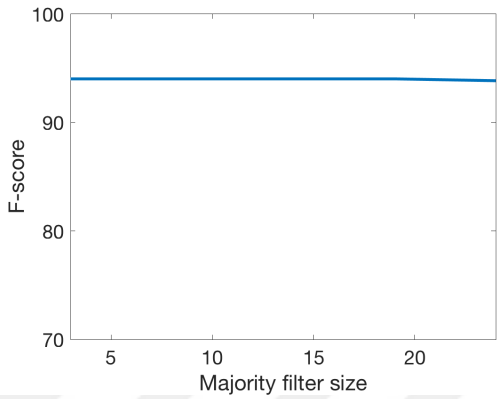


(b)

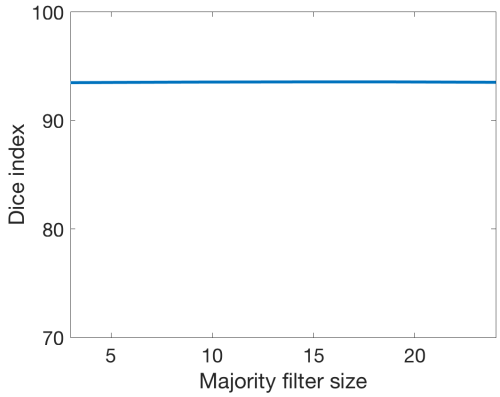


(c)

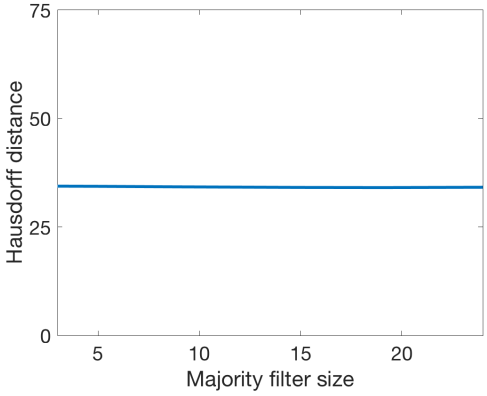
Figure 5.6: Test set F-scores, object-level Dice indices, and object-level Hausdorff distances as a function of the area threshold A_{thr}.



(a)



(b)



(c)

Figure 5.7: Test set F-scores, object-level Dice indices, and object-level Hausdorff distances as a function of the majority filter size f_{size}.

Chapter 6

Conclusion

This thesis demonstrates that a multi-stage fully convolutional model, trained adaptively to learn what to attend in images in a supervised manner, can produce better results than the existing approaches on the gland instance segmentation task in histopathological images. In contrast with the existing approaches, the proposed *AttentionBoost* model does not rely on manual and external identifications of the attention beforehand, and uses purely supervised training from fully labeled images to learn appropriate attention. To achieve this, for the first time, a new loss adjustment mechanism which uses adaptive boosting for a dense prediction model is introduced. In this mechanism, attention of each stage is modulated by separately adjusting loss contribution weights for each pixel prediction according to the errors of the previous stages on that pixel. Results on the test images show that, unlike the existing approaches with specific attention to correct boundary pixel predictions, *AttentionBoost* accounts for different types of mistake (including the boundary prediction) in the images, providing more accurate segmentation results.

The contributions of this thesis are threefold:

- It proposes the first use of an adaptive boosting approach in a dense prediction model. Although adaptive boosting is widely used for improving the

predictions of classification methods (both traditional classifiers and CNN classifiers), it is not incorporated with a dense prediction model before.

- It presents a multi-stage architecture end-to-end trainable with the proposed attention learning mechanism.
- With experiments on the gland instance segmentation task, it shows that learning the attention directly on images gives better results than using predefined attention for the model training, by reducing different types of hard-to-learn mistakes.

The framework proposed in this thesis produces the final segmentation for an image by obtaining the average of probability maps produced by all stages for that image, then applying a simple seed-controlled region growing algorithm on this map. Although we have not considered different ways to combine the output probability maps from stages, this approach could be enhanced with more advanced methods producing better overall results. For example, a supervised classification model such as neural networks can be used as a combination method. It can be considered as one future research direction of this thesis.

Additionally, even if the proposed model is investigated on gland instance segmentation task, it could also be used directly for other instance segmentation problems requiring binary classification. Likewise, another future research direction is extending this approach to other problem settings requiring multi-class classification. For this, some modifications should be made on the computation procedures of the proposed loss adjustment mechanism (in Equations 3.2 and 3.3).

We have used a simplistic FCN as a base model in the multi-stage architecture. However, the attention learning mechanism of the *AttentionBoost* model also provides the potential for the enhancement of more sophisticated model architectures. These would require advance GPU computing, but end-to-end training of many stages containing bigger models may provide benefits on difficult tasks. This is another future research direction.

Bibliography

[1] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, p. 26286, 2016.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241, Springer, 2015.

[4] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images," *Medical Image Analysis*, vol. 36, pp. 135–146, 2017.

[5] Y. Xu, Y. Li, Y. Wang, M. Liu, Y. Fan, M. Lai, I. Eric, and C. Chang, "Gland instance segmentation using deep multichannel neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 12, pp. 2901–2912, 2017.

[6] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010.

[7] K. Li, B. Hariharan, and J. Malik, "Iterative instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3659–3667, 2016.

[8] W. Shen, B. Wang, Y. Jiang, Y. Wang, and A. Yuille, "Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2391–2400, 2017.

[9] S. Gidaris and N. Komodakis, "Detect, replace, refine: Deep structured prediction for pixel wise labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5248–5257, 2017.

[10] A. Romero, M. Drozdzal, A. Erraqabi, S. Jégou, and Y. Bengio, "Image segmentation by iterative inference from conditional score estimation," *arXiv preprint arXiv:1705.07450*, 2017.

[11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[12] Y. Xu, Y. Li, M. Liu, Y. Wang, M. Lai, I. Eric, and C. Chang, "Gland instance segmentation by deep multichannel side supervision," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 496–504, Springer, 2016.

[13] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[15] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248, Springer, 2017.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

[17] H. Schwenk and Y. Bengio, "Boosting neural networks," *Neural Computation*, vol. 12, no. 8, pp. 1869–1887, 2000.

[18] D. Medera and S. Babinec, "Incremental learning of convolutional neural networks," in *IJCCI*, pp. 547–550, 2009.

[19] Y. Gao, W. Rong, Y. Shen, and Z. Xiong, "Convolutional neural network based sentiment analysis using adaboost combination," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 1333–1338, IEEE, 2016.

[20] L. Wang, B. Zhang, J. Han, L. Shen, and C.-s. Qian, "Robust object representation by boosting-like deep learning architecture," *Signal Processing: Image Communication*, vol. 47, pp. 490–499, 2016.

[21] S. Han, Z. Meng, A.-S. Khan, and Y. Tong, "Incremental boosting convolutional neural network for facial action unit recognition," in *Advances in Neural Information Processing Systems*, pp. 109–117, 2016.

[22] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a Cancer Journal for Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[23] M. Fleming, S. Ravula, S. F. Tatishchev, and H. L. Wang, "Colorectal carcinoma: pathologic aspects," *Journal of Gastrointestinal Oncology*, vol. 3, no. 3, p. 153, 2012.

[24] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, "Hematoxylin and eosin staining of tissue and cell sections," *Cold Spring Harbor Protocols*, vol. 2008, no. 5, pp. pdb–prot4986, 2008.

[25] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[26] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[27] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, pp. 818–833, Springer, 2014.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

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

[30] 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.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[33] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[34] H.-S. Wu, R. Xu, N. Harpaz, D. Burstein, and J. Gil, "Segmentation of intestinal gland images with iterative region growing," *Journal of Microscopy*, vol. 220, no. 3, pp. 190–204, 2005.

[35] A. Banwari, N. Sengar, M. K. Dutta, and C. M. Travieso, "Automated segmentation of colon gland using histology images," in *2016 Ninth International Conference on Contemporary Computing (IC3)*, pp. 1–5, IEEE, 2016.

[36] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Medical Image Analysis*, vol. 14, no. 1, pp. 1–12, 2010.

[37] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 11, pp. 2366–2378, 2015.

[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[39] W. Li, S. Manivannan, S. Akbar, J. Zhang, E. Trucco, and S. J. McKenna, "Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pp. 1405–1408, IEEE, 2016.

[40] P. Kainz, M. Pfeiffer, and M. Urschler, "Semantic segmentation of colon glands with deep convolutional neural networks and total variation segmentation," *arXiv preprint arXiv:1511.06919*, 2015.

[41] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, 2016.

[42] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical Image Analysis*, vol. 52, pp. 199–211, 2019.

[43] A. BenTaieb, J. Kawahara, and G. Hamarneh, "Multi-loss convolutional networks for gland analysis in microscopy," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 642–645, IEEE, 2016.

[44] Z. Yan, X. Yang, and K.-T. T. Cheng, "A deep model with shape-preserving loss for gland instance segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 138–146, Springer, 2018.

[45] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, pp. 179–187, Springer, 2016.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[47] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[48] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.