Machine learning system using Deep Convolutional Neural Networks to help in the assessment of colorectal surgery

Pham Tan Hung (A0112086N)

(B.Eng. (Hons.), Biomedical Engineering, NUS)

A THESIS SUBMITTED FOR THE DEGREE OF MASTER OF COMPUTING

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2019

Supervisor:

Associate Professor LEOW Wee Kheng

Examiners:

Associate Professor Terence SIM

Assistant Professor LEE Gim Hee

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

_____

Pham Tan Hung (A0112086N)

April 2019

Acknowledgments

I would like to thank my supervisor A/Prof. Leow Wee Kheng for his advice and support as well as opening the opportunity for me to do research on this thesis. In addition, I would like to thank Dr. Chong Choon Seng, from the National University Hospital, for his medical advices. I would also like to thank my family and my fiancé, who always give me full support in this journey.

Abstracts

Colorectal cancer is one of the most prevalent types of cancer in both incident rate and mortality rate. To improve treatment outcome and reduce complications, quantifying the structural characteristics of the mesorectum, colorectum, the pelvic and the boundary containing internal organs is essential.

Intra and inter-observer variability are not small in these problems, and hence an absolute ground truth is not available. However, with enough trainers preparing the data, the label will calibrate toward the ground truth. In this study, a machine learning system utilizing deep convolutional neural networks is proposed that can improve when there are more trainers and data. This system can fully automatically analyze data. Initial results for such system also show that it is very robust, consistent and efficient.

The system will be able to aid clinicians in diagnosing and screening, thus ensure higher quality treatment. Moreover, such a system will enable a larger population-based study where there can be hundreds to thousands of patients, and it is not possible for clinicians to prepare data manually.

Keywords: segmentation, landmark detection, deep learning, convolutional neural networks, learning system, fully automated, MRI, mesorectum, colorectum, pelvic dimension, colorectal cancer.

Contents

List of Tables

List of Figures

1 Introductions

1.1 Motivations

Cancer is on the rise, in term of both incident rate and mortality rate. For colorectal cancer, it was third in incidence (10.2%) out of 18.2 million new cases and second in mortality (9.2%) out of 9.6 million death in both sexes worldwide in 2018 alone (Bray et al., 2018). Most of the time it is detected at the later stages (III and IV), which severely decreases survival rate (Lee, Chew, Chow, Zheng, & Ho, 2015; Siegel et al., 2017).

Even though the mortality rate can be reduced with early and regular screening (Schreuders et al., 2015), success in treatment is one of the most crucial factors for survival rate. For rectal cancer surgery, total mesorectal excision (TME) introduced by Heald et al (Heald, Husband, & Ryall, 1982) in 1982 is the gold standard (Penna, Cunningham, & Hompes, 2017). This surgery removes the mesorectum completely (Heald et al., 1982) and has superiority over others in term of cure rates and recurrence rate (MacFarlane, Ryall, & Heald, 1993). Mesorectum refers to the region of fat that embraces the lateral and posterior sides of the retroperitoneal rectum (Diop et al., 2003). It contains all the lymphatics that may harbor cancer cells, and this is the rationale for removing the entire intact envelope within the TME plane (Torkzad, Hansson, Lindholm, Martling, & Blomqvist, 2007).

Chapter 1. Introductions

Mesorectum volume and measurement surrounding it (volume ratio with the organ chamber volume, pelvic dimension, etc.) have been shown to have impacts on the type of surgery chosen and the immediate as well as long term outcomes of the patients (Allen, Gada, & Blunt, 2007; Ishida et al., 2013; Tayyab, Razack, Sharma, Gunn, & Hartley, 2015; Torkzad & Blomqvist, 2005; Torkzad et al., 2007). For the type of surgery, there are keyhole (laparoscopic) TME (LTME) and open TME (OTME) (Vennix et al., 2014). It has been shown that LTME has speedier recovery rate post treatment with fewer complications and smaller cosmetic impact on patients (Vennix et al., 2014). However, mesorectum volume and pelvic dimension play a role in the difficulty of LTME (Allen et al., 2007; Ishida et al., 2013).

It is worth noticing that all of those mesorectum and pelvic measurements can be extracted in magnetic resonance imaging (MRI) scans, meaning clinicians can give good prediction non-invasively. Currently, almost all study concerning the mesorectum utilizes manual segmentation with the help of experts. This is a time consuming and costly procedure.

Upon realizing the importance of the mesorectum volume and other parameters around it, this thesis aims to propose a learning computer system that in time can provide reliable fully-automated parameter extraction of the meso-colorectal scans and hopefully will assist the diagnosis of clinicians in the future. The selling point of the system is in its ability to learn and grow, with every additional training case that a doctor

provides, the system gets closer to human performance. Moreover, once the system is implemented, there is no need for a computer expert to maintain the system fulltime as the training of the system can be done by anyone.

This first version of the system will consist of two main functions. The first one is to segment the mesorectum and the colorectum in order to calculate the necessary volume or area. The second function is to find landmarks to calculate the pelvic dimensions.

1.2 Thesis objectives

The objective of this thesis is to train and implement the first version of a computer system that can (1) segment the colorectal and mesorectal regions (Figure 1.1) and (2) detect six important landmarks to assist in the assessment of colorectal surgery (Figure 1.2, Figure 1.3). For tissue segmentation problem, the objective is: to segment and output 3 binary masks corresponding with 3 classes (background, mesorectum and colorectum). For landmark localization problem, the objective is: to detect six landmarks using segmentation approach.

Moreover, it is also essential to show that such a system can grow with more and more training samples. The main assumption is that the region of interest is visible in the input images and hence cases of irrelevant images are not addressed.

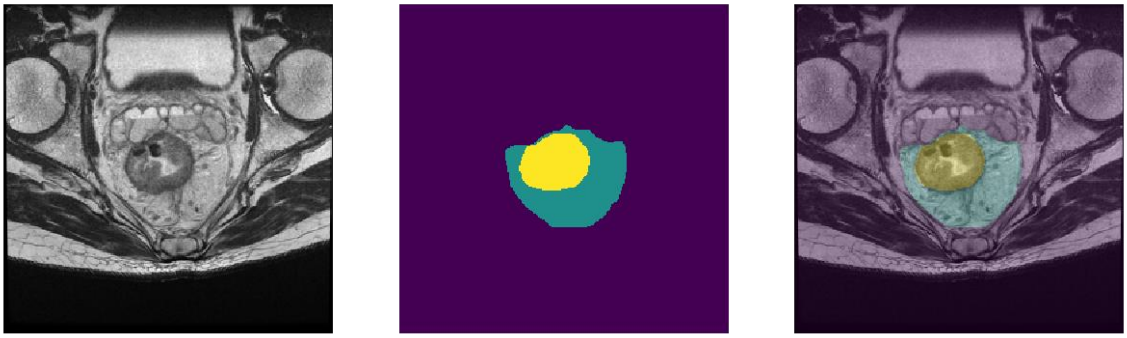


Figure 1.1. Training objective for tissue segmentation problem. Left: Input, middle: output, right: overlay between input and output

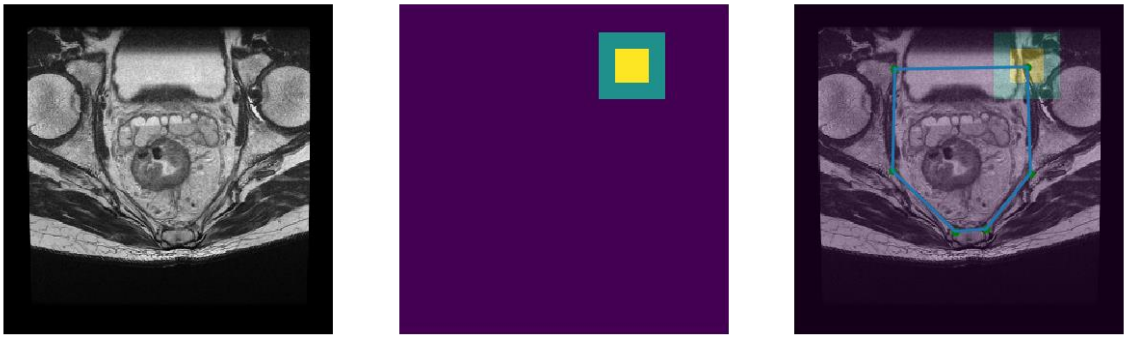


Figure 1.2. Training objective for top right landmark for landmark detection/localization problem. Left: Input, middle: output, right: overlay between input and output

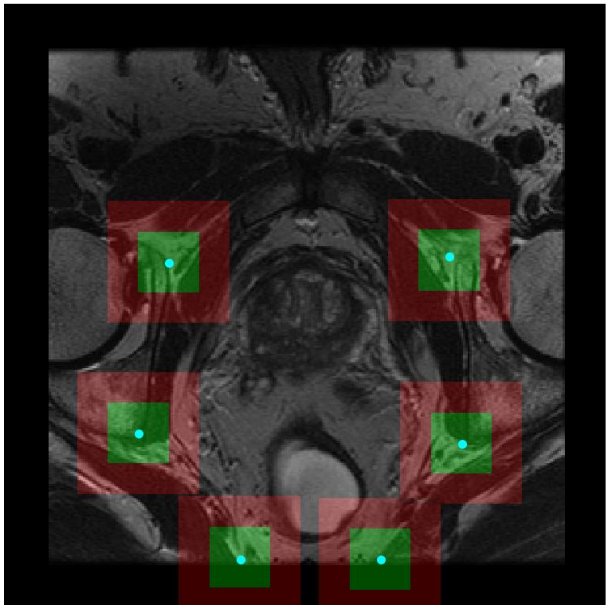


Figure 1.3. The locations of all six landmarks

2 Related works

2.1 Region-based segmentation

Methods in this approach find regions with similar pixel characteristics and boundaries can be obtained as a result of segmentation. These methods can be supervised or unsupervised.

For unsupervised methods there can be clustering-based methods (Celeux, Forbes, & Peyrard, 2003; Coleman & Andrews, 1979; Yang, Hu, Lin, & Lin, 2002; Yao, Duan, Li, & Wang, 2013), region growing methods (Adams & Bischof, 1994; Fan, Yau, Elmagarmid, & Aref, 2001) or graph-based methods that look for region homogeneity (Boykov & Funka-Lea, 2006; Boykov & Jolly, 2000; Felzenszwalb & Huttenlocher, 2004). For clustering-based methods, the image is transformed into a feature space, and then clustering rules are applied to form groups (or clusters) in this space. Clustering-based methods also consider region homogeneity. The rules or features used are mostly local and hence are very sensitive to changes. For region growing methods, image is segmented into various regions that initially started from seeds. The seeds grow by pixels' characteristics, and it can be stopped by prior knowledge rules. For graph-based methods, the image is treated as a graph G = {V,E} where V defines the nodes (or pixels) and E defines the edges connecting the nodes. The segmentation is done by partition the graph by maximizing or

minimizing certain aspects within the partition. Graphs are also usually used with Markov theory such as Markov random field (MRF) or conditional random field (CRF) in order to build contextual models (Lafferty, McCallum, & Pereira, 2001).

For supervised methods, there can be pixel-level annotation and image-level annotation. For pixel-level annotation, many state-of-the-art models utilize MRF or CRF to find the most probable label for each pixel of an image. To label a pixel, we need a unary potential for each pixel and a pairwise potential for pairs of neighborhood pixels. To obtain the unary potential, we need to train a classifier based on the strong annotation provided. The classifier can be random forest (Schroff, Criminisi, & Zisserman, 2008), SVM (Furey et al., 2000) or convolutional neural network (Krizhevsky, Sutskever, & Hinton, 2012). The pairwise potential is defined over a region surrounding the pixel of interest. After that, we can perform maximum a posterior inference. For image level annotation, one image will only have one label, oppose to a strong annotation meaning every pixel is labelled. One can use a form of expectation-maximization (EM) algorithm (Duygulu, Barnard, de Freitas, & Forsyth, 2002; Verbeek & Triggs, 2007) or multi-instance learning (MIL) which include MI - support vector machine (MI-SVM) (Andrews, Tsochantaridis, & Hofmann, 2003), in order to give the attention to the region that have classify the image. Other methods include label propagation(Liu et al., 2009) and bounding box approaches (Xia, Domokos, Dong, Cheong, & Yan, 2013). However, these

approaches based on image-level labels provide very low-quality segmentation. To improve quality, semi-supervised approaches are proposed wherein between image-level labels, we can use a few strongly annotated images. This method when coupled into an EM-based algorithm improves the performance significantly (Papandreou, Chen, Murphy, & Yuille, 2015).

2.2 Boundary-based segmentation

Methods in this approach find region boundaries that differentiate between regions and hence, boundaries are obtained directly while regions are obtained as a result of that process.

Some of the notable methods in this approach are gradient based (Delogu, Fantacci, Kasae, & Retico, 2007; Geets, Lee, Bol, Lonneux, & Grégoire, 2007), snakes (active contour) (Chan & Vese, 2001; Kass, Witkin, & Terzopoulos, 1988; C. Xu & Prince, 1998), level-set (Leventon, Faugeras, Grimson, & Wells, 2000; Li et al., 2011), watershed (Beucher, 1992; Hill, Canagarajah, & Bull, 2003), etc. Gradient based method detects boundaries based on discontinuity detection. Active contour method, also known as snakes, detects boundaries using energy minimizing, deformable spline. Snakes requires prior knowledge of the target shape or interactions with the user. Level-set method evolves a surface until some stop criteria (such as edges) to obtain the contour of a region. Watershed method treat image like a topographic map with gradients of the image is

the topographic surface. The contours are then the ridges of the image that separate the drainage regions.

2.3 Hybrid approaches segmentation

These approaches work on both regions and boundaries at the same time.

Graph cut methods (Boykov & Funka-Lea, 2006; Boykov & Jolly, 2000; Felzenszwalb & Huttenlocher, 2004) can be considered a hybrid between region-based and boundary-based segmentation since the graph works on both nodes and edges.

Active appearance model (Chen, Udupa, Bagci, Zhuge, & Yao, 2012; Mitchell et al., 2002) is a method that deforms shape and appearance of a statistical model to match target image. It uses appearance estimation and target image's difference to optimize.

There are many approaches and methods being developed for semantic image segmentation and mentioned above are just some of the notable ones.

2.4 Landmark localization

There are many landmark localization methods developed for different problems such as facial landmark localization (Zhanpeng Zhang, Luo, Loy, & Tang, 2014), pose detection (Ramanan & Zhu, 2012), and skull landmark localization (El-Feghi, Sid-Ahmed, & Ahmadi, 2004; Richtsmeier, Paik, Elfert, Cole, & Dahlman, 1995).

Notably for the current problem is the work of Tompson et al (Tompson, Jain, LeCun, & Bregler, 2014) where a hybrid network of convolutional network and MRF is proposed. This approach provides heatmap images for the landmark. Expanding on this work, Payer et al (Payer, Štern, Bischof, & Urschler, 2019) encode pseudo-probability of a landmark being located at a certain position using such heatmap. The landmark location is then at the coordinate where the heatmap has the highest value (Payer et al., 2019).

3 Methods

Three neural network architectures will be explored in this thesis, namely U-net, Full Resolution Residual Network (FRRN) and Full Resolution Residual U-net (FRRUnet). They are all composed of convolutional layer.

First, section 3.1 will introduce the proposed approach. An ensemble model composed of the three networks will also be explored. Convolutional neural networks will be discussed in section 3.2 while the three architectures will be discussed in section 3.3 to 3.5. Ensemble model will be explored in section 3.6 and neural network training will be described in section 3.7.

3.1 Proposed approach

Based on the related work of landmark localization, I proposed an approach where the heatmap turns into a segmentation map, with the landmark lies exactly in the middle of the region.

Hence, segmentation will be used for both problems. Moreover, due to the rise of deep convolutional neural networks, it would be very interesting to study the segmentation results of such approach with very limited data and how it improves when given more data. Hence, convolutional neural networks are used throughout this study.

There are two objectives for the learning system. The first one is to segment the mesorectum and the colorectum. The second one is to localize the

landmarks to measure pelvic dimensions and assess the volume of the entire boundary surrounding the internal organs.

Details on segmentation approach using deep convolutional neural networks will be explored in subsequent sections in this chapter.

3.2 Convolutional neural networks

Convolutional neural networks are neural networks that utilize convolutional layer as building blocks. Convolutional layer will consist of K number of filters that will transform an input with dimensions H*W*1 into H'*W'*K (Figure 3.1). The filter size is 3 by 3 (Figure 3.1). In neural networks, a convolutional layer (Krizhevsky et al., 2012; LeCun & Bengio, 1995; LeCun, Kavukcuoglu, & Farabet, 2010; Simonyan, Vedaldi, & Zisserman, 2013) is defined by filter/kernel size, number of filters/kernels, padding size and stride size. For a 2D image, 2D filters are normally used. But it is also possible to use a 3D filter for a multi-channel image (e.g.: color image with RGB channels). It is nonetheless possible to do 2D convolution on a 3D volume since each filter is extended through the whole depth of the input. Since the thesis deals with grayscale images (one channel), 2D filters are used.

For the current problem, every convolutional layer will have:

- K filter
- F*F filter size
- Stride S

- Padding P

An image with input dimensions W*H*D will have output dimensions W'*H'*K with:

$$W' = \frac{W - F + 2P}{S} + 1$$

$$H' = \frac{H - F + 2P}{S} + 1$$

However, in the current thesis, we used F = 3, S = 1, P = 1, effectively cancels out everything beside W and H, hence making W' = W and H' = H.

The weights of each filter are shared for the entire image, constraining them to do the same operation on different parts of the image. However, different filters in one layer will have different weights, enabling multiple feature extraction at each location (LeCun & Bengio, 1995). Moreover, since each convolutional layer has multiple filters, there is a risk that all filters will converge to the same set of weights, making them effectively equal to one filter. Details on how to avoid this will be described and explained in 3.6.2 Training procedures.

Image is downsampled 2 times in order to compress the representation and extract important features. However, the number of filters is doubled after every downsampling stage to prevent drastically reducing the number of neurons. Overall, there will be 4-6 downsampling and upsampling stages with the current input image size of 256 by 256. More than that then the latent space dimensions will be too small.

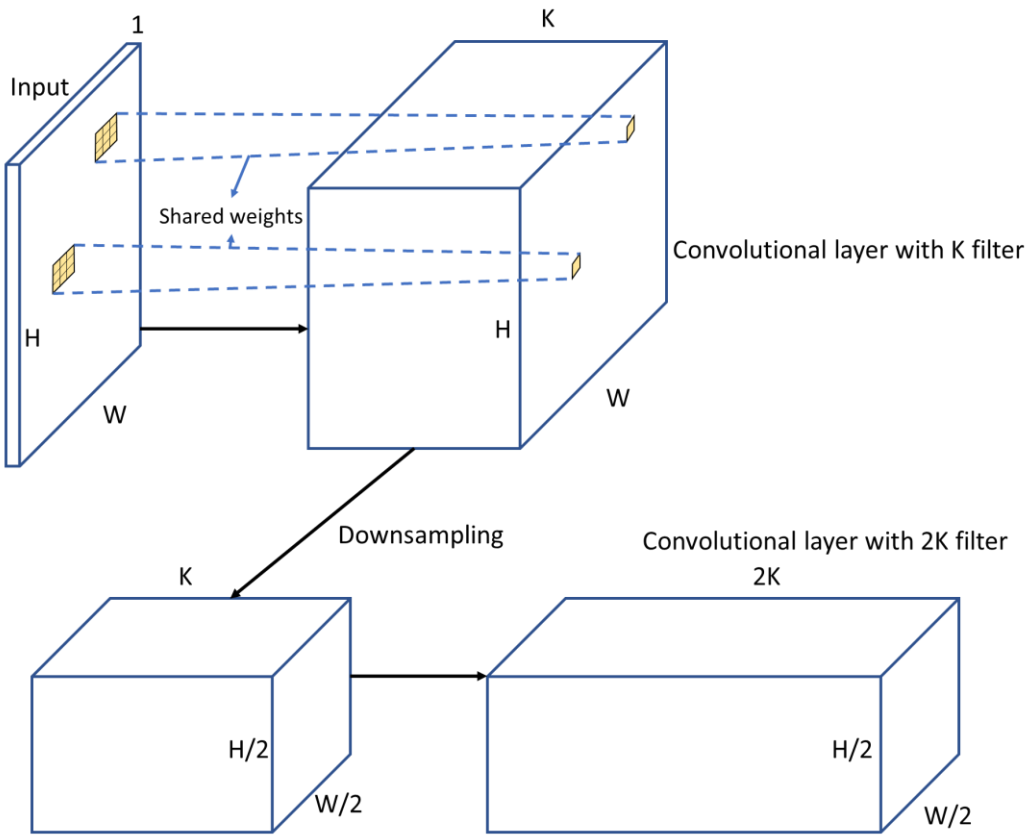


Figure 3.1. Overview of convolutional neural network

This type of convolutional layer is adopted throughout all three architecture.

3.3 U-net

U-net architecture (Ronneberger, Fischer, & Brox, 2015) is given in figure 3.2. U-net follows closely the architecture of an autoencoder, in which the inputs are being down-sampled in order to learn more higher abstract features and then up-sampled again to output the desired objective, in this case, it is the segmented masks of the same size as the inputs. Before each step of down-sampling and up-sampling, there are two convolutional

layers, and each layer is followed by an activation function. All activation functions used before the last layer are rectified linear unit (ReLU) (Dahl, Sainath, & Hinton, 2013; Nair & Hinton, 2010; B. Xu, Wang, Chen, & Li, 2015): $f(x) = \max(0, x)$. These activation functions help the network to learn the non-linearity in the data. The last layer activation is a Softmax function(Bishop, 2006): $f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$.

Moreover, the number of filters for each convolutional layer (number in the bracket after layer name in figure 3.2) doubles for each down-sampling step and halves for each up-sampling step. The starting number of filters is 32 (same for FRRN and FRRUNet). Through trials and errors, these numbers of filters are found to be producing good prediction accuracy while still being able to be loaded into memory. After each down-sampling (MaxPooling) layer, the dimensions of the image become H'/2 by W'/2, where H' and W' are the dimensions of the previous layer. After each up-sampling layer, the dimensions of the image become 2*H' by 2*W', where H' and W' are the dimensions of the previous layer. Refer to each architecture figure for the layer dimensions written in *Italic*. All convolutional layer used padding so that the output dimensions stayed the same as the input dimensions. Only after maxpooling and upsampling that the image H and W dimensions change, while only convolutional layers will change the third dimension (number of channels/filters) (beside reshaping and concatenation).

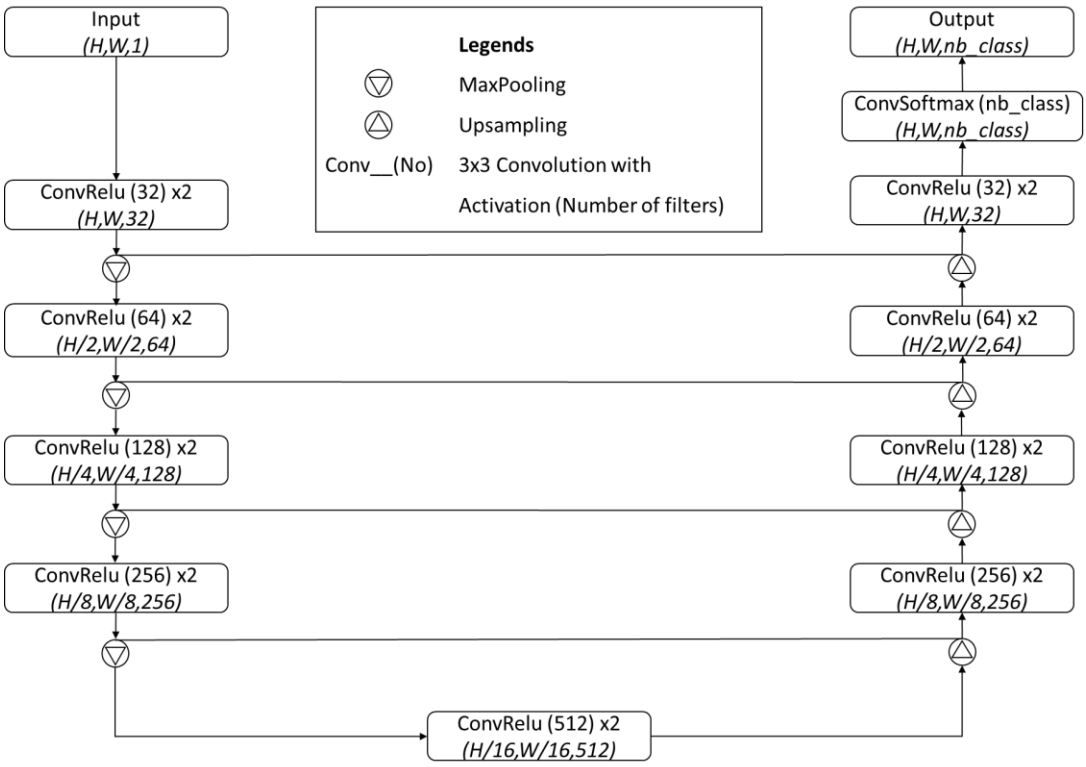


Figure 3.2. U-net Architecture

The most significant difference of U-net compares to an autoencoder is the skip connections that connect each down-sampling step with the corresponding up-sampling step (Ronneberger et al., 2015). These skip connections provide more spatial data that is lost in down-sampling steps, hence making the final predictions more robust and smoother.

Total number of parameters in U-net is 7,846,147 and all parameters are trainable.

3.4 Full resolution residual network (FRRN)

FRRN architecture (Pohlen, Hermans, Mathias, & Leibe, 2017) is given in figure 3.3. The overall architecture is still the same as U-net, downsampling

the inputs to extract features and then upsampling to get the desired outputs. However, instead of using skip connections to ferry forward spatial data, FRRN keeps a full resolution highway (z-stream in figure 3.3). The information on this highway is modified after each down-sampling and up-sampling step and also before those steps, information from the highway is pulled and concatenate with the y-stream (autoencoder stream). This kind of highway is shown to help the network learn faster with less data.

Moreover, the building blocks for FRRN are not a simple convolutional layer like U-net but are consist of full resolution residual unit (FRRU), residual unit (RU) and convolutional batch normalization (Ioffe & Szegedy, 2015) block (ConvBN block). Details on those building blocks are provided in figure 3.3. Basically, an N by N will denote filter size, and the number in the bracket will denote the number of filters for each layer. Since this network needs to maintain a full resolution highway, it uses more memory than U-net with the advantages of faster learning. However, since it uses more memory, the number of filters after each MaxPooling and UpSampling layer cannot be doubled and halved respectively. Hence the number of filters only increases by 16 and 32 for the first two stages and the last four stages respectively. Same goes for UpSampling stages.

The activation functions used before the last layer are leaky ReLUs(B. Xu et al., 2015):

$$f(x) = \begin{cases} \alpha x, & x < 0 \\ x, & x \geq 0 \end{cases}$$

Where α is a small constant to keep the gradient alive even if x is negative. For this thesis, α is chosen to be 0.3. Batch normalization finds the mean and variance of each mini batch and then scale and shift the values (subtract the mean and divide by the standard deviation) (Ioffe & Szegedy, 2015).

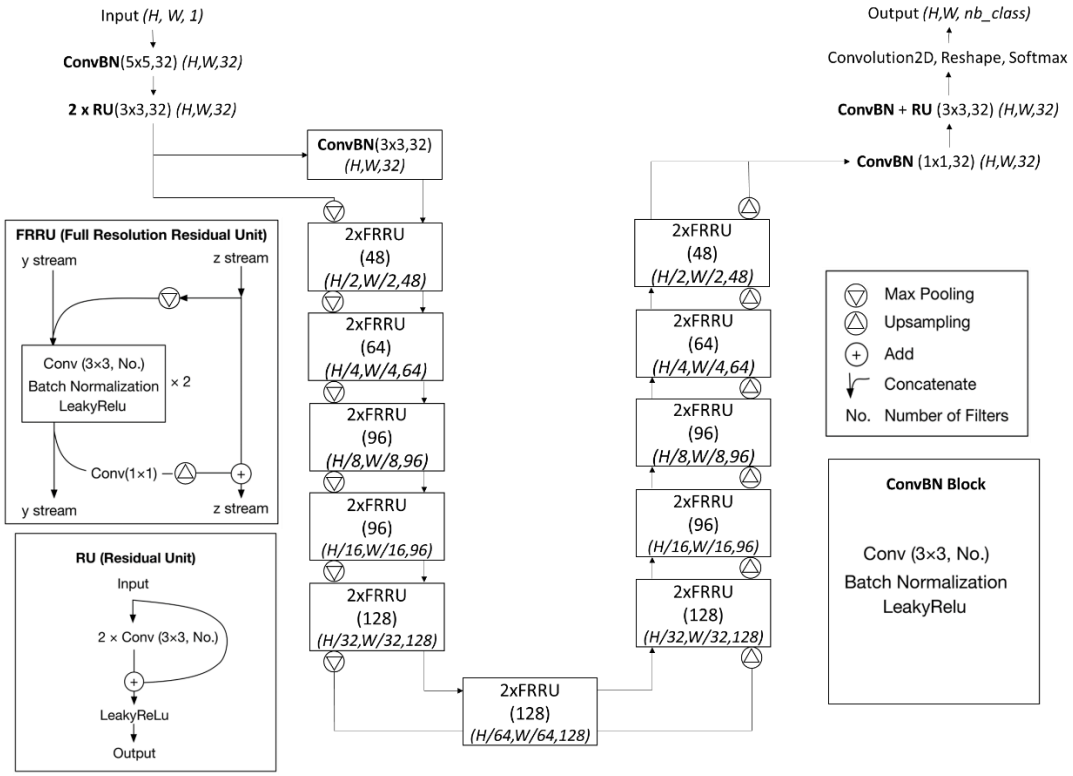


Figure 3.3. FRRU architecture

Total number of parameters in FRRN is 4,263,843. There are 4,255,843 trainable parameters and 8,000 non-trainable parameters from batch normalization layers.

3.5 Full resolution residual U-net (FRRUnet)

This architecture called FRRUnet is a combination of U-net and FRRN (Figure 3.4). The proposed model utilized U-net (Figure 3.2) skip connections to connect the first two down-sampling (Encoding) and up-sampling (Decoding) blocks. However, the skip connections for FRRUnet have a feature extraction residual unit (RU) before concatenating with the up-sampling stream. This is to extract meaningful data instead of just bring all the information forward. The rest of the model followed closely that of FRRN (Figure 3.3). The number of filters for each layer follows the same rule as FRRN because this network also has the disadvantage of using more memory, even though it uses less than FRRN because the highway is at half resolution.

Total number of parameters in FRRUnet is 3,973,571. There are 3,966,371 trainable parameters and 7,200 non-trainable parameters from batch normalization layers.

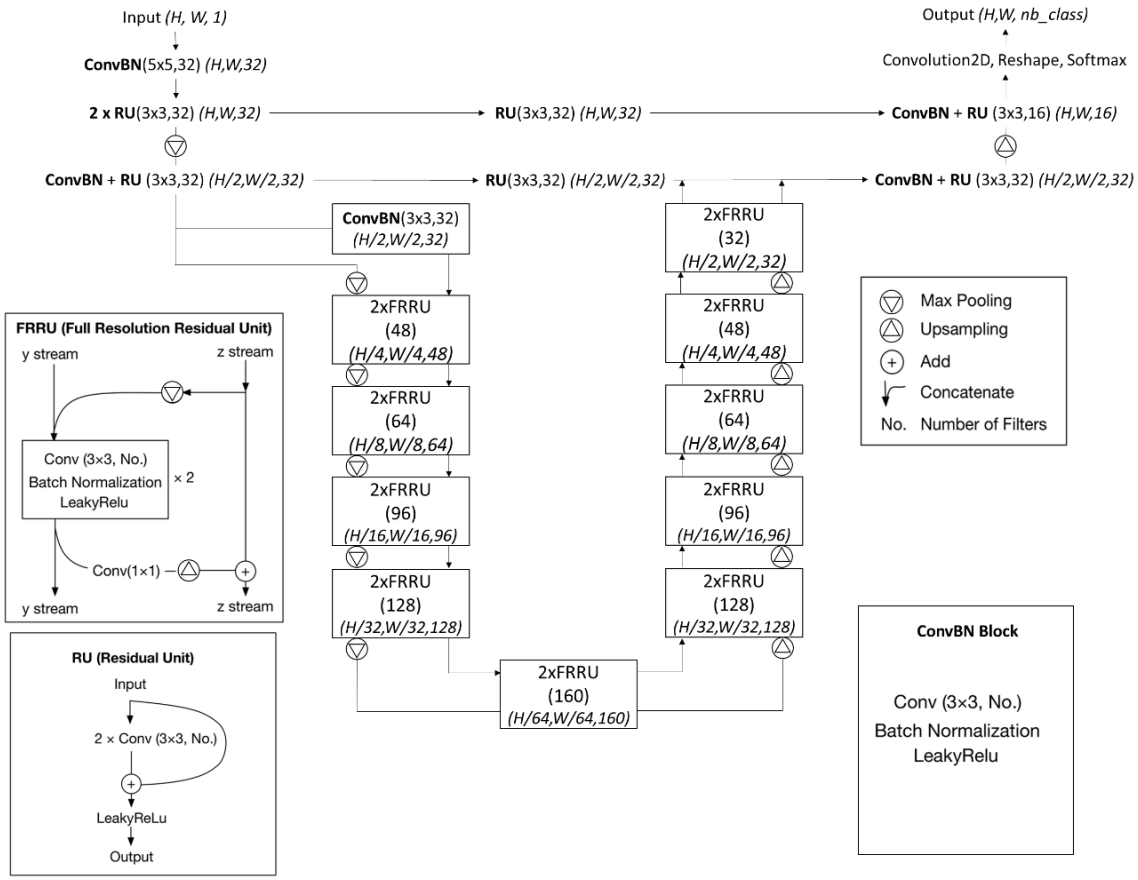


Figure 3.4. FRRUnet architecture

3.6 Ensemble model

For segmentation problem, an ensemble method (Figure 3.5)(Hansen & Salamon, 1990; Perrone & Cooper, 1992) was introduced. This method learned how to combine the three base models (U-net, FRRN, FRRUnet) in a non-linear manner using an MLP. The base models are frozen and non-trainable. All outputs for each pixel from all models were concatenated along the last dimension and used as input to a 100 hidden-unit MLP. The output dimensions of each model are H*W*3, and the concatenated dimensions are H*W*9. The weights/parameters that connected the input

to the hidden layer and to the output were shared between all pixels. The output dimension of the MLP was the same as the base models. The objective of this problem is shown in figure 3.6.

To create diversity in learning performance and bias, the number of downsampling and upsampling was different for each base model. For U-net, it had 4 stages (Up-Down sampling). FRRN had 6 stages. FRRUnet had 2 U-net stages and 4 FRRN stages (6 total). It is good to have different performance and bias because when the predictions are combined in ensemble method there will be more information at different image location. If the bias is the same for all network, they will fail at the same locations and even combining them will not help to recover from that.

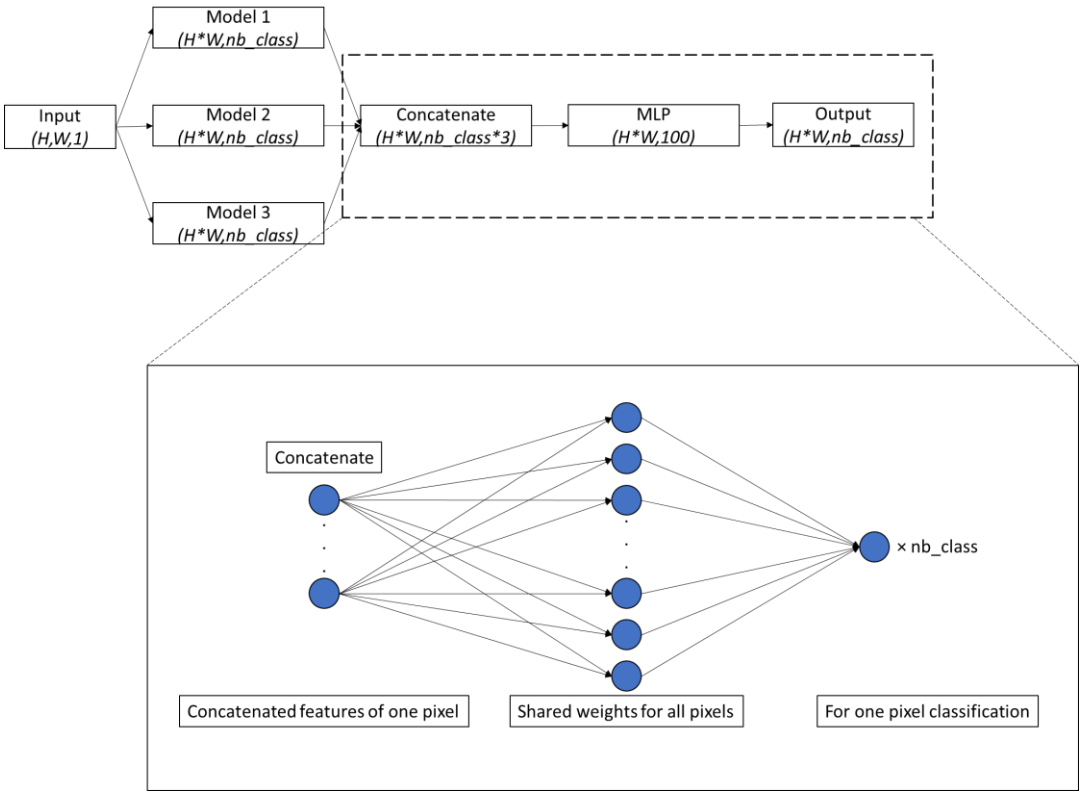


Figure 3.5. Ensemble architecture

Total number of parameters in ensemble model is 16,084,864. There are 1,303 trainable parameters and 16,083,561 non-trainable parameters from frozen base models.

3.7 Neural networks training

3.7.1 Training objectives

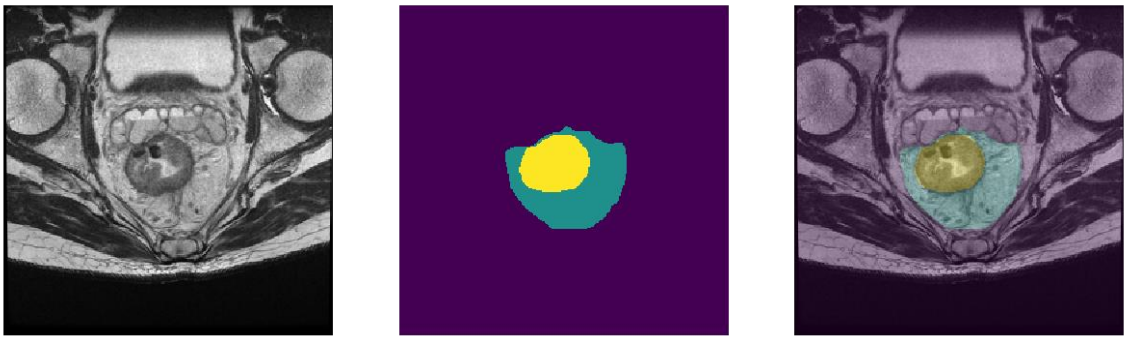


Figure 3.6. Training objective for tissue segmentation problem. Left: Input, middle: output, right: overlay between input and output

For tissues segmentation, the input and output dimension height (H) and width (W) are 256 and 256 respectively. For landmark localization, H and W are 296 and 296 respectively due to a padding of size 20. Both tissue segmentation and landmark localization have three classes. The dimensions are fixed, but since Dicom images are of square ratio, the input can be easily resized without losing over structural properties to the suitable dimensions.

For tissue segmentation problem, the objective is quite clear: to segment and output 3 binary masks corresponding with 3 classes (background, mesorectum and colorectum).

For landmark detection problem, the network needed to identify 6 landmarks and training an ensemble model would require a lot of time and resources (each model needed to train 4 times, 3 base models and 1 ensemble model then time 6 landmarks, total of 24 models). Hence, only FRRUnet was used for this problem. The 6 landmarks are used to approximate the pelvic dimensions as well as to draw a boundary of the chamber that contains the internal organs and they are simply named: top left, top right, middle left, middle right, bottom left, bottom right – based on their locations on the scan. Traditionally to identify a Cartesian coordinate, a regression model is used since the outputs are continuous. However, due to the fact that medical images are limited in quantity, it is very tough to train a robust regression model. Hence a segmentation approach was proposed (Figure 3.7). The point landmark of interest was expanded into a small box, termed as the Focus Region, with the point located exactly centrally. Moreover, to reduce class imbalance, a larger region was selected surrounding the Focus Region, called Attention Region. This region helped direct the network attention to the region of interest.

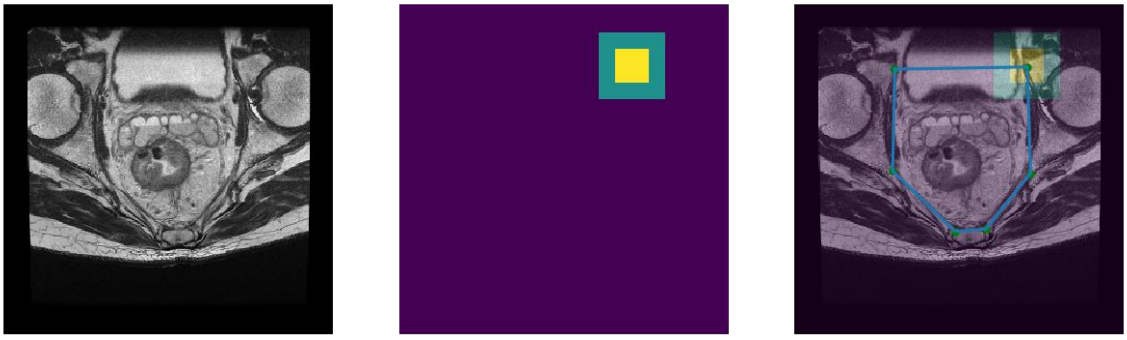


Figure 3.7. Training objective for top right landmark for landmark detection/localization problem. Left: Input, middle: output, right: overlay between input and output

A general definition of a regression approach is as follow:

$$h: R^{(H \times W)} \Rightarrow R^2$$

where the dimensions of the image are H and W pixels for height and width respectively. The outputs for such an approach are two real values for x and y location. During training, there will only be two feedbacks (back propagation (Horikawa, Furuhashi, & Uchikawa, 1992)) from x and y for the network to optimize.

Whereas for a segmentation approach proposed, the whole process was divided into two steps:

$$h: R^{(H \times W)} \Rightarrow R^{(H \times W \times 3)}$$

$$c: R^{(H \times W \times 3)} \Rightarrow R^2$$

The first step would predict 3 binary masks of the same size as the input. These masks corresponded to 3 classes: Focus Region, Attention Region

and Background. This would make $H \times W$ feedbacks to the network during training since every pixel needed to be classified. The second step was to calculate the center point of the region to get the x and y location of the landmarks. The formula to calculate the center of the regions is:

Centroid = $\left(\frac{\mu_{1,0}}{\mu_{0,0}}, \frac{\mu_{0,1}}{\mu_{0,0}}\right)$where $\mu_{m,n} = \sum_{x=0}^{w} \sum_{y=0}^{h} (x - c_x)^m \left(y - c_y\right)^n M(x, y)$

Where M is the image moment (Mukundan & Ramakrishnan, 1998).

Since all neural networks were segmentation (or pixelwise classification) networks, their final activation layers were all softmax (Bishop, 2006). Softmax calculates the probabilities distribution of each pixel over all possible classes. The formula for softmax is:

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \text{ (Bishop, 2006)}$$

3.7.2 Training procedures

To train a neural network, an optimizer and a loss function are necessary. Loss function gives feedback about how the prediction is different from the ground truth (target/objective). These losses are then back propagated throughout the network, and the network's parameters are tuned using a variation of gradient descent called Adam (Kingma & Ba, 2014).

For a multi-class classification like the current problems, categorical cross entropy (Zhilu Zhang & Sabuncu, 2018) was used as the loss function. The formula is:

$$l = -\sum_i^C y_i \log(p_i)$$ (Zhilu Zhang & Sabuncu, 2018)

where y_i is the ground truth and p_i is the prediction after a softmax activation for each class i in C.

Adam (Kingma & Ba, 2014) was used as optimizer. The formula for updating the parameters is as follow:

$$m = \beta_1 m + (1 - \beta_1) \times d_x$$

$$v = \beta_2 v + (1 - \beta_2) \times {d_x}^2$$

$$x += \frac{-\text{learning\_rate} \times m}{\sqrt{v} + \varepsilon}$$

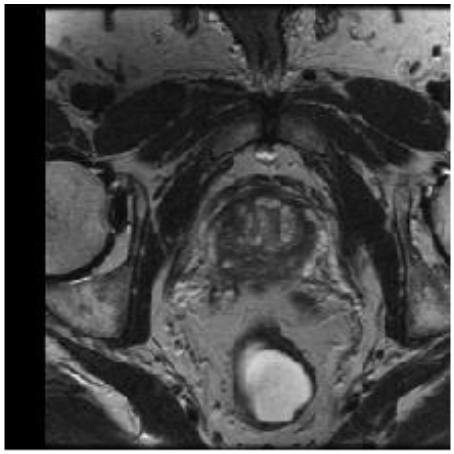
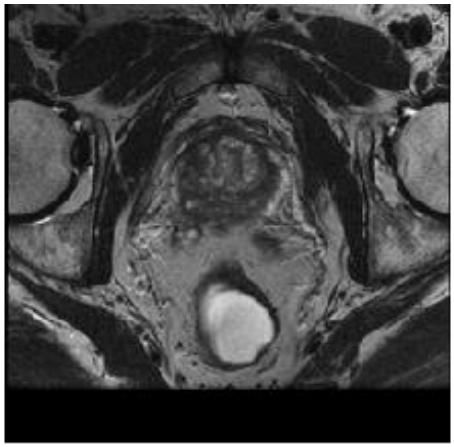
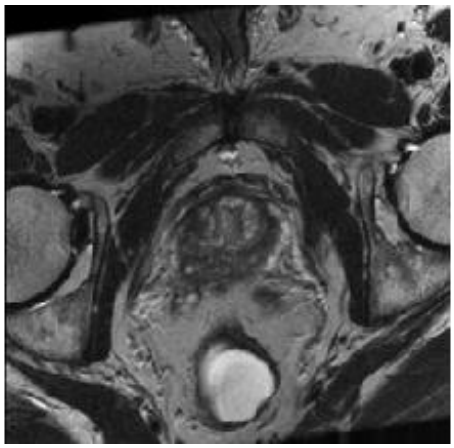
where x is the updated parameter, d_x is the gradient and $\beta_1, \beta_2, \varepsilon$ are hyperparameters. After trials and errors, the learning rate was chosen to be 0.00005 for the base models and 0.0001 for the ensemble model. The hyperparameters were left as recommended from the paper, β_1: 0.9, β_2: 0.999, ε: $1e-7$.

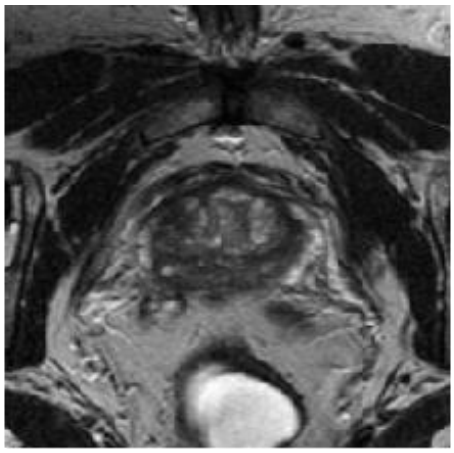
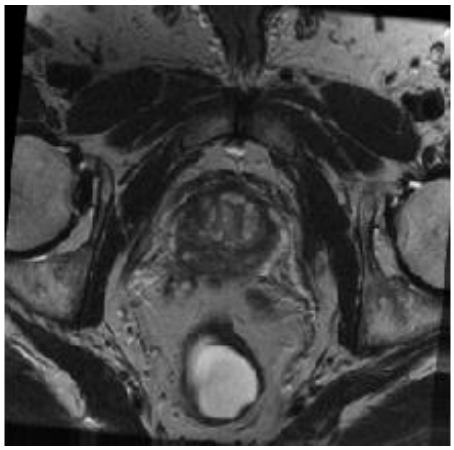
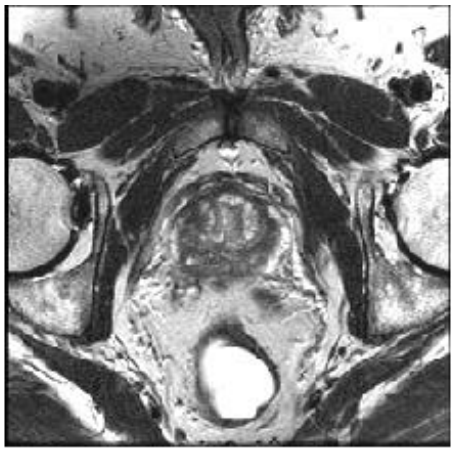
To address the problem of limited data, augmentation was used. Details are in table 3.1.

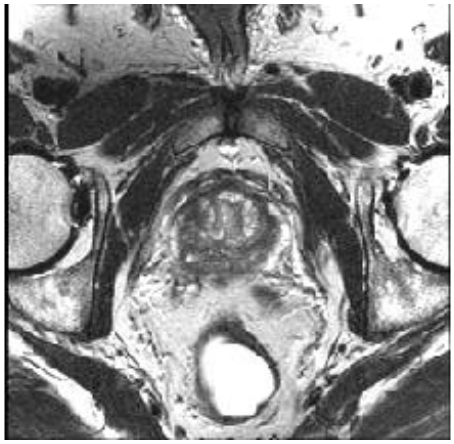
For tissue segmentation, additional random horizontal flip was applied.

Checkpoint was also used to save the best model so far on validation set instead of only saving the latest model.

Table 3.1. Augmentation used in training

Deformation	Range	Example
Width shift	0.05	
Height shift	0.15	
Shear	0.05	

Zoom	0.9 - 1.05	
Rotation	8 degree	
Brightness	[-15, +15]	

Contrast	0.8 – 1.5	

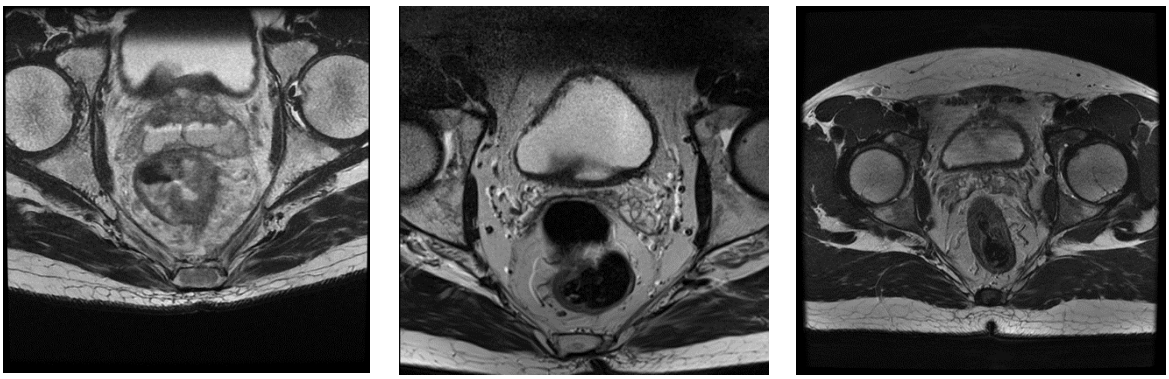
Moreover, to prevent overfitting, dropout (Hinton, Krizhevsky, Sutskever, & Srivastva, 2014) was used frequently after big layers. Dropout means dropping out weights/parameters in the neural network with a predefined probability. In this thesis, the chosen probability is 0.5. Dropout also helps to prevent the filter from converging into the same set of weights since for every training epoch, half of the filters' weights are set to 0 on random, forcing them to learn different features.

3.7.3 Model retraining

One of the advantages of a machine learning system is that it can easily be retrained. Once additional training data are labelled, one can proceed to retrain the neural network models using python scripts. It is up to the user to change the hyper-parameters such as learning rate, checkpoint, batch size, number of epochs, and even the model architectures.

4 Experiments and Discussions

4.1 Data preparation and preprocessing



A B C

Figure 4.1. Example scans from the extended dataset. A: scan from NUH. B: scan from PROMISE 12 dataset. C: scan from 108 Military Hospital.

4.1.1 Dataset

Data collected from National University Hospital (NUH) (Singapore), 108 Military Hospital (Vietnam) and Prostate MRI Segmentation (PROMISE12)(Litjens et al., 2014) dataset were used for model training and testing. For data collected from hospitals, it was in Dicom format, and pyDicom Python package was used to extract the data. Axial scans were used. PROMISE12 dataset was already extracted. For PROMISE12 dataset, slices that contain the mesorectum section were manually selected. Since this dataset focuses on prostate cancer, only a small portion of the slices contained the mesorectum. Even though the angle of the axial scans in PROMISE was different from colorectal scans, the overall visual was similar and the crucial tissues were present (Figure 4.1).

FIJI (Schindelin et al., 2012) was used to prepare the ground truth for tissue segmentation, and MATLAB (MathWorks, Inc) was used to prepare the ground truth for landmark detection.

4.1.2 Tissue segmentation

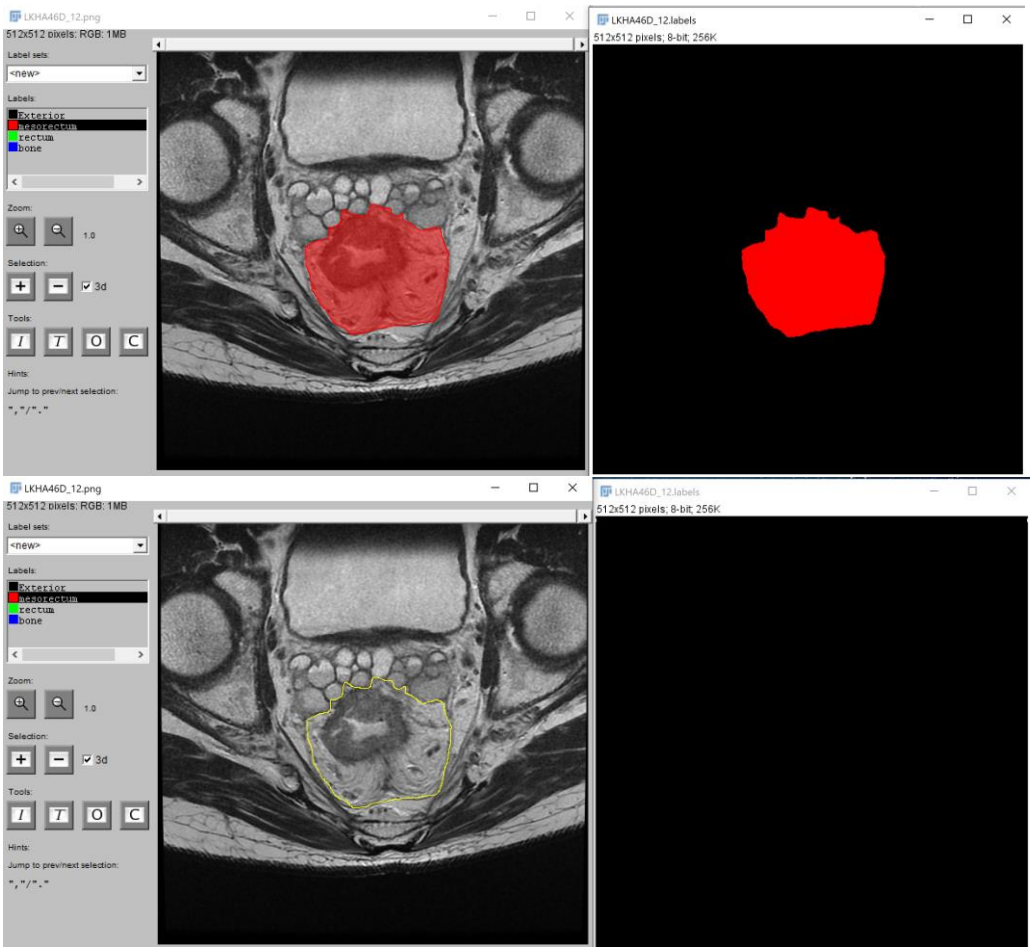


Figure 4.2. Data preparation procedure for tissue segmentation problem

For tissue segmentation, a small macro was written to utilize FIJI and its Segmentation plugin. The user only needs to open the image and run the macro in FIJI and start segmenting like Figure 4.2. The user will need to select the corresponding tissue of interest before segmenting. The '+' and '-' signs will add or subtract the selection into or from the tissue

respectively in the label image shown on the right. The tools available for region selection are from FIJI toolbars such as polygon, eclipse and freehand etc. Once finished, the label will be saved to the training folder.

4.1.3 Landmark location

For landmark detection labelling, a MATLAB script was written to assist human/trainer/user in quick data preparation. The user will mark the 6 landmarks, press enter once to show the confirmation points in red squares (Figure 4.3). If the user is not satisfied, they can proceed to re-labelling the location. Once the user is satisfied with the marking, press enter without selecting any point on the confirmation screen will bring the next image.

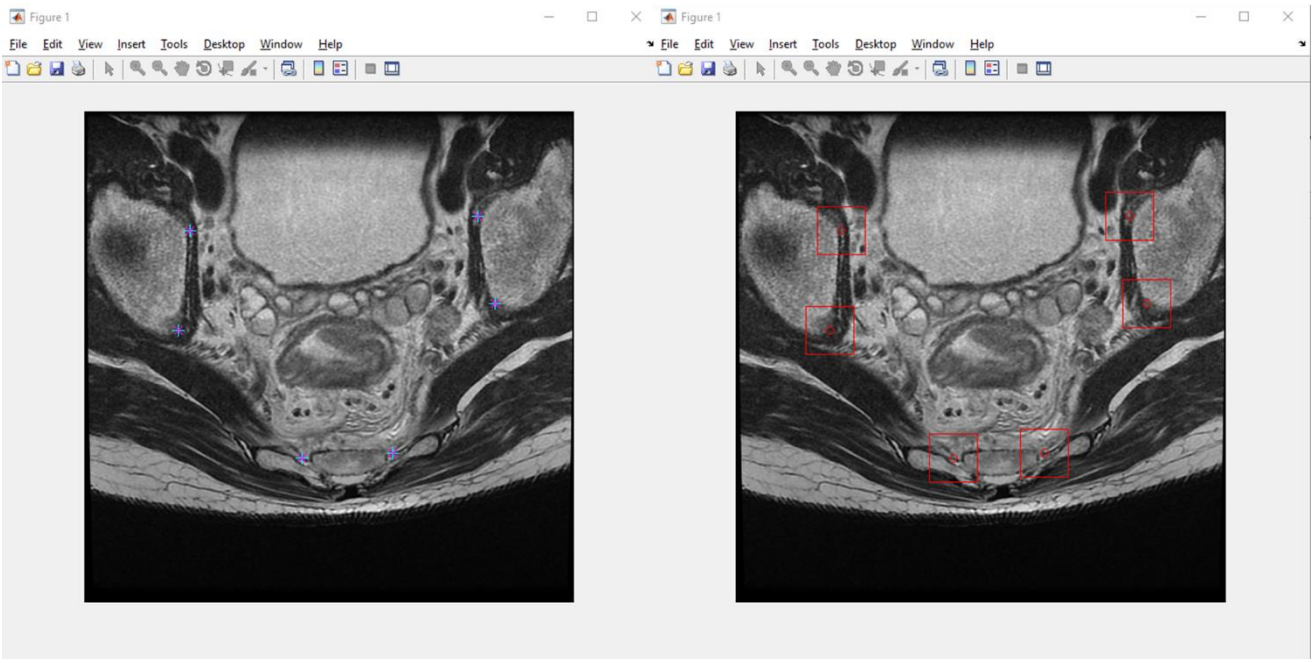


Figure 4.3. Data preparation procedure for landmark detection/ localization problem.

4.2 Test procedures

Training and testing are executed in my PC with the following specifications. The graphical processing unit (GPU) was NVIDIA GTX 1080 (NVIDIA Corporation), and the central processing unit (CPU) was an Intel CPU core i7-7700 (Intel Corporation) at clock speed 3.6 GHz. Single threading was used for CPU workload. The neural network programs are implemented using Keras (Chollet, 2017) and Tensorflow (Abadi et al., 2016) in Python 3.

The main test results will only be the NUH test dataset. The pixel resolution for this dataset with image height and width of 256 by 256 is 0.703125 mm. All Dicom images are of square ratio and hence resizing images to 256 by 256 will not affect the tissue structures. All measurements reported will be in real-world scale.

4.2.1 Tissue segmentation

To show that those additional data were helpful even though they were not from the exact problem, the segmentation model was trained with the original NUH data only that consist of 4 training samples, 2 validation samples and 8 testing samples; and with the extended data set (NUH + 108 Military Hospital + PROMISE 12) that consist of 102 training samples, 27 validation samples and two set of testing data. The first testing set was the original 8 testing samples (NUH dataset), and the second set is the extended one with 18 testing samples (NUH + 108 Military Hospital).

The main scores usually used to assess a segmentation against a ground truth are Dice score(Crum, Camara, & Hill, 2006; Milletari, Navab, & Ahmadi, 2016), sensitivity (true positive) and specificity (true negative) (Altman & Bland, 1994):

$$Dice\ score\ = \frac{2 \times |A \cap B|}{2 \times |A \cap B| + |B \backslash A| + |A \backslash B|}$$

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

$$Specificity\ = \frac{TN}{TN\ +\ FP}$$

, where TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

Those scores were used to assess tissue segmentation robustness and accuracy. Dice score is the main benchmark for accuracy as it is based on overlapping index between the prediction and ground truth of each tissue and hence is not susceptible to class imbalance.

4.2.2 Augmentation ablation study

Beside additional data, image augmentation plays an important role in preventing overfitting as well as increasing model accuracy. Beside the basic augmentations such as image shifting, zooming and flipping, this thesis also utilized contrast and brightness shifting as well as rotation. To

show the usefulness of contrast-brightness shifting and rotation in training the model, a comparison between no augmentation, only contrast-brightness shift, only rotation and all augmentation will be studied. The model used is FRRUnet and the score for comparison is dice score. Extended dataset was used for training and testing.

4.2.3 Landmark localization

Once the extended dataset was proven to be useful, the full dataset (NUH + 108 Military Hospital + PROMISE 12) was used for training (102 samples), validating (27 samples) and testing (18 samples) of the landmark detection network.

Intraclass correlation coefficient (ICC) is a statistical measurement used extensively in the study of reliability between observers (inter-observer) and within single observer (intra-observer)(Koo & Li, 2016). It is chosen to be the primary benchmark to measure how well the machine learns to mimic human trainer as well as to compare between approaches. ICC values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9 and greater than 0.9 are considered poor, moderate, good and excellent reliability respectively (Koo & Li, 2016). There are many types of ICC, the one chosen in this experiment is two-way, single score, absolute agreement ICC based on guideline paper (Koo & Li, 2016).

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)},$$

where: MS_R = mean square for rows

MS_E = mean square for error

MS_C = mean square for columns

n = number of subjects

k = number of raters/measurements

Since the location of a single landmark is two dimensional (x and y), ICC is calculated for each dimension and average to obtain the final reliability score. Average distances between predicted location and human marked location are also used to give a more well-rounded perspective.

4.2.4 Machine learning system efficiency

Training time and testing (inference) time are recorded across multiple runs to show the efficiency of the system.

The training time for FRRN, U-net, FRRUnet and Ensemble model are 260ms/step, 106ms/step, 202ms/step and 76ms/step respectively for a batch size of 3 per step. All models are trained for 10 steps per epoch with 2000 epochs. The total training time for FRRN, U-net, FRRUnet and Ensemble model are 86 minutes, 35 minutes, 67 minutes and 25 minutes respectively. Ensemble model training is much quicker than other models even though it is bigger is due to almost all parameters are frozen and it only has around one thousand trainable parameters.

The inference time for Ensemble and FRRUnet model was recorded as they were the final models used for result analysis. For Ensemble model, the

prediction time per scan is 0.05 ± 0.0036 seconds. If the whole extended test dataset (18 samples) were loaded on memory and prediction was run on the whole set, the prediction time reduce to 0.036 ± 0.00024 seconds per scan. For FRRUnet model, the prediction time per scan is 0.03 ± 0.0025 seconds. If the whole extended test dataset (18 samples) were loaded on memory and prediction was run on the whole set, the prediction time reduce to 0.017 ± 0.00034 seconds per scan.

4.3 Tissue segmentation results and discussions

The results of the scores are presented in table 4.1, 4.2 and 4.3 for three experiments. Scores in **bold** are the highest in the current column of that tissue while scores with an underline are the lowest. However, for many cases, the performances are so close that there is no real single highest or lowest score.

Dice score shows the overall segmentation quality and is one of the main benchmarks used to compare between models and experiments.

This section contains 3 experiments to show the machine robustness and potential in colorectal MRI image tissue segmentation. For the first test, the models were only trained and tested on NUH dataset. The results are shown in table 4.1. Overall, ensemble model managed to capture more general patterns and the final Dice scores for mesorectum and colorectum segmentation.

Table 4.1. Test results from training and testing on original NUH dataset

	Dice	Sensitivity	Specificity
Mesorectum			
U-net	75.41 ± 7.05	**87.74** ± 4.01	98.10 ± 0.84
FRRU	79.38 ± 5.70	84.86 ± 6.37	98.81 ± 0.31
FRRUnet	75.91 ± 8.99	71.96 ± 12.93	99.35 ± 0.20
Ensemble	**79.98** ± 7.17	76.37 ± 11.41	**99.44** ± 0.23
Colorectum			
U-net	86.31 ± 4.87	85.54 ± 4.90	99.63 ± 0.31
FRRU	85.32 ± 7.98	83.67 ± 9.65	**99.64** ± 0.28
FRRUnet	80.86 ± 11.28	81.97 ± 13.00	99.22 ± 1.42
Ensemble	**88.11** ± 6.10	**87.79** ± 7.22	99.61 ± 0.65

The second test (table 4.2) is to show the learning system potential in the future, by expanding the training set while keeping the test set to be the same. The dice score increases nearly 8% for mesorectum segmentation which is one of the main objectives of the thesis. The colorectum segmentation has mixed results, dice score still has an average increase of 2.4%, with a sharp increase in FRRU model and a slight decrease in U-net and Ensemble. Overall, expanding the dataset results in a net increase in the performance of the networks.

Table 4.2. Test results from training on extended dataset and testing on original NUH dataset

	Dice	Sensitivity	Specificity
Mesorectum			
U-net	81.26 ± 5.21	88.40 ± 5.48	98.81 ± 0.39
FRRU	85.09 ± 3.69	84.39 ± 8.15	**99.45** ± 0.35
FRRUnet	83.17 ± 5.49	**91.44** ± 3.12	98.85 ± 0.47
Ensemble	**85.69** ± 4.31	91.38 ± 4.19	99.11 ± 0.40
Colorectum			
U-net	85.67 ± 5.88	76.66 ± 7.97	**99.95** ± 0.05
FRRU	**92.09** ± 3.13	**88.11** ± 5.71	99.91 ± 0.07
FRRUnet	82.82 ± 7.32	73.05 ± 11.32	99.93 ± 0.07
Ensemble	87.95 ± 5.56	80.63 ± 8.42	99.94 ± 0.06

The slight decrease in performance in some models might be because the networks now need to learn a larger dataset with possibly more complex patterns and in order to have a better generalization, it loses some specific patterns in the original dataset. Hence for the third experiment (table 4.3), the test set is also expanded. Indeed, both the mesorectum and colorectum got an increase of around 6.5% compared to the original dataset.

Table 4.3. Test results from training and testing on extended dataset

	Dice	Sensitivity	Specificity
Mesorectum			
U-net	80.99 ± 10.79	**89.24** ± 7.86	99.02 ± 0.75
FRRU	82.30 ± 5.38	80.20 ± 8.78	**99.59** ± 0.30
FRRUnet	82.53 ± 7.42	87.19 ± 6.47	99.24 ± 0.50
Ensemble	**85.08** ± 5.70	88.49 ± 5.96	99.40 ± 0.40
Colorectum			
U-net	90.76 ± 6.17	85.65 ± 10.04	**99.95** ± 0.04
FRRU	**92.39** ± 2.99	**92.18** ± 6.66	99.86 ± 0.09
FRRUnet	88.46 ± 7.81	86.49 ± 14.25	99.85 ± 0.12
Ensemble	91.49 ± 5.30	89.71 ± 9.99	99.89 ± 0.08

In conclusion, these experiments show that the models can learn to segment important tissues with high accuracy as well as sensitivity and specificity. Moreover, the potential of a learning system is clearly proven as given more data, the model improves significantly.

4.4 Augmentation study results and discussions

Results are shown in table 4.4. It clearly shows the importance of data augmentation in training deep neural networks, especially in the case of limited data.

Table 4.4. Test results of various augmentation configuration from training and testing on extended dataset

Configuration	Mesorectum	Colorectum	Background
No augmentation	59.38±10.2	71.60±26.58	99.08±0.34
Brightness + contrast	63.45±12.94	75.81±8.92	99.09±0.47
Rotation	73.59±7.08	77.08±21.27	99.40±0.26
All augmentation	82.53±7.42	88.46±7.81	99.66±0.13

4.5 Landmark localization results and discussions

No landmark in regression approach obtains excellent reliability and only one location has good reliability while half of the landmarks in the proposed segmentation approach receive excellent scores and one receive a good score. For the rest of the landmark in the segmentation approach, the ICC is still larger than the one in regression approach by several folds. The details of the results are shown in table 4.5.

Table 4.5. ICC score and average distance of predicted landmarks against human markers

	Top left	Top right	Middle left	Middle right	Bottom left	Bottom right
Regression						
ICC x	0.806	0.481	0.236	0.132	0.034	-0.109
ICC y	0.867	0.795	0.731	-0.123	-0.001	0.009
Average ICC	0.837	0.638	0.484	0.005	0.016	-0.050
Average Distance (mm)	6.858 ± 3.699	11.901 ± 4.205	9.999 ± 7.367	25.546 ± 12.756	20.540 ± 5.781	27.385 ± 7.798
Segmentation						
ICC x	0.934	0.805	0.947	0.974	0.298	0.222
ICC y	0.884	0.931	0.906	0.960	0.836	0.556
Average ICC	0.909	0.868	0.927	0.967	0.567	0.389
Average Distance (mm)	5.617 ± 3.373	5.726 ± 1.963	4.022 ± 2.173	3.339 ± 1.398	4.354 ± 1.737	4.686 ± 1.936

To explain why for the proposed segmentation approach the ICCs for some landmarks are low but the average distances are still close to human marks, we look at ICC in more details. ICC considers both the distance between two raters as well as the trend in marking of both raters. If all top right x

coordinates in the segmentation approach are subtracted by 10, meaning the trend between samples is still the same, the ICC drop to 0.384 from 0.805. This explains why bottom point predictions have low ICC but shorter distances to human markers. Visualization for x coordinate of top right and bottom right landmarks further proves the point. The top right x coordinate predictions follow the human marks not just in term of distance but also the bias (the overall trend). The bottom right x coordinate predictions might be closer to human's but the overall trend between human and machine looks random. For regression method, if the distance is too great, no matter how close the trend is between two raters, the ICC will still be very low.

Overall, this proves that the proposed segmentation approach is superior by a large margin in all landmark detection, especially in the case of limited training data. Moreover, it can be concluded that the proposed segmentation approach can detect the landmark reliably.

4.6 Learning system analysis

The whole result analysis process for a single volume of 30 scans takes only a few seconds to finish with commercially available hardware. Moreover, once the system is trained, its prediction is deterministic, meaning with the same input it will always produce the same output. So, the consistency is high. Finally, in term of robustness and accuracy, it is shown in the first two sections of this chapter that the machine can learn from human with high

sensitivity and specificity. Nonetheless, with more training data, the machine proved to be improving significantly.

Normally, a deep neural network needs huge amount of data (can come up to millions of samples) in order to predict with high accuracy. However, there are two main reasons why the current neural network system needs much fewer data. Firstly, the system adopts a segmentation approach, which is very rich in information for labelled data. Every pixel has a label, not just one label for the entire image like normal classification task. Thus, making back propagation information rich and hence making the network learn faster. Secondly, there are much fewer data variations in medical images. Unlike normal real life images in which the main object can have a lot of variations (e.g.: for a cat it can be black, brown or it can be sitting, standing, hiding etc.), medical images always have a standard operation during the image capture process and the tissue general location and orientation should be similar between images.

The visualization of the segmentation and landmark localization results can be found in the appendix. One example of not up to standard segmentation is shown here in figure 4.5. This might be due to the tissues' boundaries are not clearly seen and the network detected the wrong boundaries.

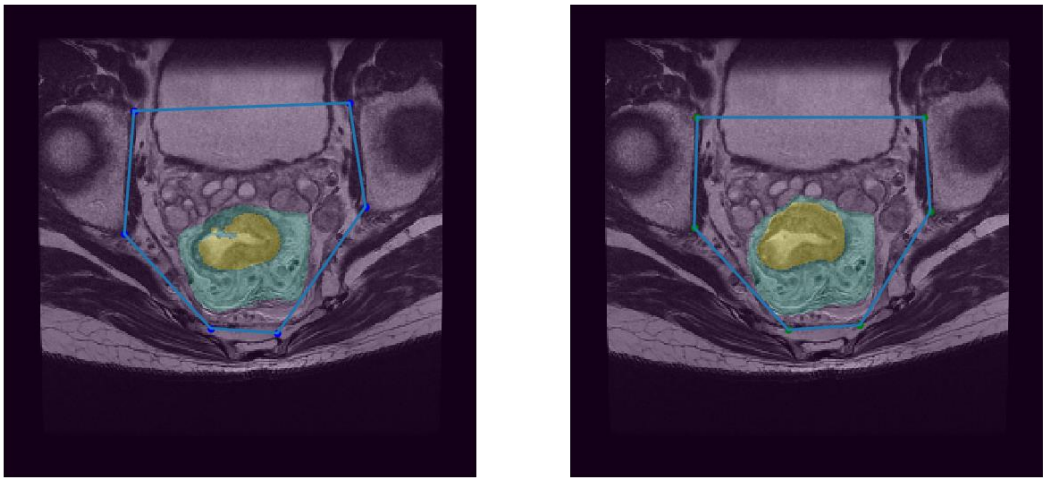


Figure 4.5. An example output from the machine (left) compare to human (right). Segmentation result here needs improvement

5 Future works and Conclusions

5.1 Future works

The current prototype of the learning system is proved to be robust, consistent and efficient. However, as of the current state, it still lacks a quality check and exception catching function. Exception catching implementation is a matter of software development and should be implemented in the future.

For prediction quality check, the preliminary algorithm is already thought out. For tissue segmentation, a simple clustering algorithm can be used because of the prior knowledge that for one axial scan, there should be only one or two clusters of mesorectum and colorectum. For landmark detection, a confidence score is proposed due to the nature of the segmentation approach. Confidence score is based on the predicted shape of the Focus region, if it follows a square with a fixed side length the confidence will be higher and vice versa.

Another technical limitation is from the landmark localization approach. Since it needs a surrounding region from the point of interest, it might not work if the point is too close to the border. The problem was faced in this study for the bottom landmarks. However, it was solved by adding a padding border with the width similar to half of the focus region side length.

The networks are deterministic with the same input; however, it was not tested against augmented or adversarial inputs. This will be a very interesting study for the future.

Another limitation is the lack of hyper-parameter tuning such as which number of filters is optimal for each layer. This is due to the search space for this problem is large while time and resources are limited. However, in the future, a thorough grid search to arrive at optimal hyper-parameter values will be appreciated.

As an exploratory study, there is an obvious lack of data and expert trainers. However, thanks to the nature of this thesis objective, a working prototype of a learning system is successfully implemented and proven to be able to improve significantly with more data and trainers. Moreover, since the point of machine learning system proposed is to allow doctors or medical experts to be able to teach the machine themselves, the system has the potential to be very close to an expert if it is put into a clinical setting usage. Hence, in the future, with more data and expert trainers, the system can be much better.

5.2 Conclusions

Advancement in screening and diagnosing is vital in the control and treatment of colorectal cancer. Analyzing and quantifying the mesorectum, colorectum, as well as pelvic structures, can lead to a more suitable surgery option as well as reduce complications and increase long term survival rate

for patients. It is true that many procedures are time-consuming with intra- and inter-observer variability. However, statistically, with more trainers (observers), the results should calibrate toward a ground truth.

Hence, a learning system has been proposed. With this system, many clinicians can train the machine together without the expert knowledge in coding. Moreover, with many expert trainers, the machine is bound to perform much better, with it being able to get very close to a ground truth is entirely possible.

Currently, the system can segment and measure the mesorectum and colorectum volume as well as pelvic dimension and quantify the overall structure of the boundary that contains the internal organs. With a modular approach, the system can do much more, as long as an explicit objective function is described, and proper training data is prepared.

Moreover, such a system like this, besides aiding clinicians, will enable large population study between the structures of mesorectum, colorectum, etc. and prognosis, risk factors and many other aspects. Many potential links and correlations can be found that is difficult before due to the tedious nature of labelling the data.

In conclusion, a learning system is introduced that can quantify many structural parameters that can help in the assessment of colorectal cancer surgery with high robustness, consistency and efficiency. Such a system is

also able to improve itself with clinicians' input and training, making it future proof.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Isard, M. (2016). *Tensorflow: A system for large-scale machine learning.* Paper presented at the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16).

Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence, 16*(6), 641-647.

Allen, S., Gada, V., & Blunt, D. (2007). Variation of mesorectal volume with abdominal fat volume in patients with rectal carcinoma: assessment with MRI. *The British journal of radiology, 80*(952), 242-247.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal, 308*(6943), 1552.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). *Support vector machines for multiple-instance learning.* Paper presented at the Advances in neural information processing systems.

Beucher, S. (1992). The watershed transformation applied to image segmentation. *SCANNING MICROSCOPY-SUPPLEMENT-*, 299-299.

Bishop, C. M. (2006). *Pattern recognition and machine learning*: springer.

Boykov, Y., & Funka-Lea, G. (2006). Graph cuts and efficient ND image segmentation. *International journal of computer vision, 70*(2), 109-131.

References

Boykov, Y., & Jolly, M.-P. (2000). *Interactive organ segmentation using graph cuts.* Paper presented at the International conference on medical image computing and computer-assisted intervention.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians, 68*(6), 394-424.

Celeux, G., Forbes, F., & Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern recognition, 36*(1), 131-144.

Chan, T. F., & Vese, L. A. (2001). Active contours without edges. *IEEE transactions on image processing, 10*(2), 266-277.

Chen, X., Udupa, J. K., Bagci, U., Zhuge, Y., & Yao, J. (2012). Medical image segmentation by combining graph cuts and oriented active appearance models. *IEEE transactions on image processing, 21*(4), 2035-2046.

Chollet, F. (2017). Keras (2015). In.

Coleman, G. B., & Andrews, H. C. (1979). Image segmentation by clustering. *Proceedings of the IEEE, 67*(5), 773-785.

Crum, W. R., Camara, O., & Hill, D. L. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging, 25*(11), 1451-1461.

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). *Improving deep neural networks for LVCSR using rectified linear units and dropout.* Paper

presented at the 2013 IEEE international conference on acoustics, speech and signal processing.

Delogu, P., Fantacci, M. E., Kasae, P., & Retico, A. (2007). Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier. *Computers in Biology and Medicine, 37*(10), 1479-1491.

Diop, M., Parratte, B., Tatu, L., Vuillier, F., Brunelle, S., & Monnier, G. (2003). " Mesorectum": the surgical value of an anatomical approach. *Surgical and Radiologic Anatomy, 25*(3-4), 290-304.

Duygulu, P., Barnard, K., de Freitas, J. F., & Forsyth, D. A. (2002). *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary.* Paper presented at the European conference on computer vision.

El-Feghi, I., Sid-Ahmed, M. A., & Ahmadi, M. (2004). Automatic localization of craniofacial landmarks for assisted cephalometry. *Pattern recognition, 37*(3), 609-621.

Fan, J., Yau, D. K., Elmagarmid, A. K., & Aref, W. G. (2001). Automatic image segmentation by integrating color-edge extraction and seeded region growing. *IEEE transactions on image processing, 10*(10), 1454-1466.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision, 59*(2), 167-181.

References

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*(10), 906-914.

Geets, X., Lee, J. A., Bol, A., Lonneux, M., & Grégoire, V. (2007). A gradient-based method for segmenting FDG-PET images: methodology and validation. *European journal of nuclear medicine and molecular imaging, 34*(9), 1427-1438.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*(10), 993-1001.

Heald, R., Husband, E., & Ryall, R. (1982). The mesorectum in rectal cancer surgery—the clue to pelvic recurrence? *British Journal of Surgery, 69*(10), 613-616.

Hill, P. R., Canagarajah, C. N., & Bull, D. R. (2003). Image segmentation using a texture gradient based watershed transform. *IEEE transactions on image processing, 12*(12), 1618-1633.

Hinton, G. E., Krizhevsky, A., Sutskever, I., & Srivastva, N. (2014). System and method for addressing overfitting in a neural network. In: Google Patents.

Horikawa, S.-I., Furuhashi, T., & Uchikawa, Y. (1992). On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm. *IEEE transactions on Neural Networks, 3*(5), 801-806.

References

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Ishida, T., Okabayashi, K., Hasegawa, H., Ishii, Y., Kikuchi, H., Seishima, R., & Kitagawa, Y. (2013). pelvic dimensions and mesorectal volume as predictors of surgical difficulty in laparoscopic surgery for rectal cancer: p094. *Colorectal Disease, 15*, 40.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision, 1*(4), 321-331.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine, 15*(2), 155-163.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in neural information processing systems.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks, 3361*(10), 1995.

References

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). *Convolutional networks and applications in vision.* Paper presented at the Proceedings of 2010 IEEE International Symposium on Circuits and Systems.

Lee, H., Chew, L., Chow, K. Y., Zheng, H., & Ho, W. (2015). Singapore Cancer Registry Annual Registry Report Trends in Cancer Incidences in Singapore 2009–2013. *National Registry of Diseases Office*.

Leventon, M. E., Faugeras, O., Grimson, W. E. L., & Wells, W. M. (2000). *Level set based segmentation with intensity and curvature priors.* Paper presented at the Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No. PR00737).

Li, C., Huang, R., Ding, Z., Gatenby, J. C., Metaxas, D. N., & Gore, J. C. (2011). A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. *IEEE transactions on image processing, 20*(7), 2007-2016.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., . . . Zhang, J. (2014). Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis, 18*(2), 359-373.

Liu, X., Cheng, B., Yan, S., Tang, J., Chua, T. S., & Jin, H. (2009). *Label to region by bi-layer sparsity priors.* Paper presented at the Proceedings of the 17th ACM international conference on Multimedia.

References

MacFarlane, J., Ryall, R., & Heald, R. (1993). Mesorectal excision for rectal cancer. *The lancet, 341*(8843), 457-460.

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). *V-net: Fully convolutional neural networks for volumetric medical image segmentation.* Paper presented at the 2016 Fourth International Conference on 3D Vision (3DV).

Mitchell, S. C., Bosch, J. G., Lelieveldt, B. P., Van der Geest, R. J., Reiber, J. H., & Sonka, M. (2002). 3-D active appearance models: segmentation of cardiac MR and ultrasound images. *IEEE transactions on medical imaging, 21*(9), 1167-1178.

Mukundan, R., & Ramakrishnan, K. (1998). *Moment functions in image analysis-theory and applications*: World Scientific.

Nair, V., & Hinton, G. E. (2010). *Rectified linear units improve restricted boltzmann machines.* Paper presented at the Proceedings of the 27th international conference on machine learning (ICML-10).

Papandreou, G., Chen, L.-C., Murphy, K. P., & Yuille, A. L. (2015). *Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Payer, C., Štern, D., Bischof, H., & Urschler, M. (2019). Integrating Spatial Configuration into Heatmap Regression Based CNNs for Landmark Localization. *Medical image analysis*.

References

Penna, M., Cunningham, C., & Hompes, R. (2017). Transanal total mesorectal excision: why, when, and how. *Clinics in colon and rectal surgery, 30*(05), 339-345.

Perrone, M. P., & Cooper, L. N. (1992). *When networks disagree: Ensemble methods for hybrid neural networks*. Retrieved from

Pohlen, T., Hermans, A., Mathias, M., & Leibe, B. (2017). Fullresolution residual networks for semantic segmentation in street scenes. *arXiv preprint*.

Ramanan, D., & Zhu, X. (2012). *Face detection, pose estimation, and landmark localization in the wild.* Paper presented at the Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Richtsmeier, J. T., Paik, C. H., Elfert, P. C., Cole, T. M., & Dahlman, H. R. (1995). Precision, repeatability, and validation of the localization of cranial landmarks using computed tomography scans. *The Cleft Palate-Craniofacial Journal, 32*(3), 217-227.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation.* Paper presented at the International Conference on Medical image computing and computer-assisted intervention.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Schmid, B. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods, 9*(7), 676.

Schreuders, E. H., Ruco, A., Rabeneck, L., Schoen, R. E., Sung, J. J., Young, G. P., & Kuipers, E. J. (2015). Colorectal cancer screening: a global overview of existing programmes. *Gut, 64*(10), 1637-1649.

Schroff, F., Criminisi, A., & Zisserman, A. (2008). *Object Class Segmentation using Random Forests.* Paper presented at the BMVC.

Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G., Barzi, A., & Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: a cancer journal for clinicians, 67*(3), 177-193.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Tayyab, M., Razack, A., Sharma, A., Gunn, J., & Hartley, J. E. (2015). Correlation of rectal tumor volumes with oncological outcomes for low rectal cancers: does tumor size matter? *Surgery today, 45*(7), 826-833.

Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). *Joint training of a convolutional network and a graphical model for human pose estimation.* Paper presented at the Advances in neural information processing systems.

Torkzad, M. R., & Blomqvist, L. (2005). The mesorectum: morphometric assessment with magnetic resonance imaging. *European radiology, 15*(6), 1184-1191.

Torkzad, M. R., Hansson, K. A., Lindholm, J., Martling, A., & Blomqvist, L. (2007). Significance of mesorectal volume in staging of rectal

cancer with magnetic resonance imaging and the assessment of involvement of the mesorectal fascia. *European radiology, 17*(7), 1694.

Vennix, S., Pelzers, L., Bouvy, N., Beets, G. L., Pierie, J. P., Wiggers, T., & Breukink, S. (2014). Laparoscopic versus open total mesorectal excision for rectal cancer. *Cochrane Database of Systematic Reviews*(4).

Verbeek, J., & Triggs, B. (2007). *Region classification with markov field aspect models.* Paper presented at the 2007 IEEE Conference on Computer Vision and Pattern Recognition.

Xia, W., Domokos, C., Dong, J., Cheong, L.-F., & Yan, S. (2013). *Semantic segmentation without annotating segments.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Xu, C., & Prince, J. L. (1998). Snakes, shapes, and gradient vector flow. *IEEE transactions on image processing, 7*(3), 359-369.

Yang, M.-S., Hu, Y.-J., Lin, K. C.-R., & Lin, C. C.-L. (2002). Segmentation techniques for tissue differentiation in MRI of ophthalmology using fuzzy clustering algorithms. *Magnetic Resonance Imaging, 20*(2), 173-179.

References

Yao, H., Duan, Q., Li, D., & Wang, J. (2013). An improved K-means clustering algorithm for fish image segmentation. *Mathematical and Computer Modelling, 58*(3-4), 790-798.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2014). *Facial landmark detection by deep multi-task learning.* Paper presented at the European conference on computer vision.

Zhang, Z., & Sabuncu, M. (2018). *Generalized cross entropy loss for training deep neural networks with noisy labels.* Paper presented at the Advances in Neural Information Processing Systems.

Appendix

Figures comparing the machine predictions (left) and human labels (right) for NUH test dataset.