11-11-2020 2:00 PM

Deep Reinforcement Learning in Medical Object Detection and Segmentation

Dong Zhang, *The University of Western Ontario*

Supervisor: Shuo Li, *The University of Western Ontario*

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

Abstract

Medical object detection and segmentation are crucial pre-processing steps in the clinical workflow for diagnosis and therapy planning. Although deep learning methods have achieved considerable performance in this field, they impose several shortcomings, such as computational limitations, sub-optimal parameter optimization, and weak generalization. Deep reinforcement learning as the newest artificial intelligence algorithm has great potential to address the limitation of traditional deep learning methods, as well as obtaining accurate detection and segmentation results. Deep reinforcement learning has a cognitive-like process to propose the area of desirable objects, thereby facilitating accurate object detection and segmentation. In this thesis, we deploy deep reinforcement learning into two challenging and representative medical object detection and segmentation tasks: 1) Sequential-Conditional Reinforcement Learning (SCRL) for vertebral body detection and segmentation by modeling the spine anatomy with deep reinforcement learning; 2) Weakly-Supervised Teacher-Student network (WSTS) for liver tumor segmentation from the non-enhanced image by transferring tumor knowledge from the enhanced image with deep reinforcement learning. The experiment indicates our methods are effective and outperform state-of-art deep learning methods. Therefore, this thesis improves object detection and segmentation accuracy and offers researchers a novel approach based on deep reinforcement learning in medical image analysis.

Keywords: Deep reinforcement learning, Medical object detection and segmentation, Vertebral body segmentation, Liver tumor segmentation, Teacher-student framework

Lay Summary

Automatic medical object detection and segmentation based on artificial intelligence as computer-assisted-diagnosis tools are significant for clinicians in the disease diagnosis and treatment planning. Medical object detection and segmentation distinguish the object of interest from the medical image, which provides clinicians with the location, shape, and size of the object, thereby assisting clinicians to make a decision. Deep learning methods have achieved considerable performance in this field by leveraging convolutional neural networks. However, as the development of deep learning, it also imposes some limitations and its accuracy in some tasks cannot meet clinical expectations. In this case, this thesis seeks to employ deep reinforcement learning to address the limitations of deep learning methods and obtain accurate medical object detection and segmentation. Particularly, this thesis deploys deep reinforcement learning in vertebral body segmentation, where the newly-proposed Sequential-Conditional Reinforcement Learning (SCRL) models the spine anatomy as a sequential decision-making process and segments vertebral bodies along the spine. In another project, this thesis deploys deep reinforcement learning into a more challenging task. Particularly, this thesis proposes the Weakly-Supervised Teacher-Student network (WSTS) to address liver tumor segmentation from the non-contrast-enhanced image. WSTS leverages deep reinforcement learning to transfer tumor spatial information for the contrast-enhanced image in the training stage, which plays as guidance to determine the liver tumor location in the non-contrast-enhanced image. The results of the above two methods outperform the results of existing deep learning methods. The success of proposed methods in medical object detection and segmentation indicates the deep reinforcement learning can be a reliable computer-assisted-diagnosis tool and benefit to clinicians.

Co-Authorship Statement

The following thesis consists of 2 manuscripts and both of them have been submitted to a peer-reviewed journal. Dong Zhang, as the first author, was a significant contributor to both studies, framework propose, validation experiment design and implementation, data analysis, and manuscript preparation. Dr. Shuo Li, as the principal investigator and supervisor, provided guidance and aided in the study conception, direction, and data acquisition. Additionally, Dr. Li was responsible for the approval and submission of the manuscripts.

Chapter 2 is an original research article entitled, "Sequential Conditional Reinforcement Learning for Automatic Vertebral Body Detection and Segmentation". This manuscript was submitted to the peer-reviewed journal *Medical Image Analysis*. This manuscript was co-authored by Dong Zhang Bo Chen, and Shuo Li.

Chapter 3 is an original research article entitled, "Weakly-Supervised Teacher-Student Network for Liver Tumor Segmentation from Non-enhanced Images". This manuscript was submitted to the peer-reviewed journal *Medical Image Analysis*. This manuscript was co-authored by Dong Zhang and Shuo Li.

Acknowledgements

I would like to thank my supervisor, Dr. Shuo Li for his support and guidance in my research and life. Thank you for teaching me research skills, including academic paper writing and scientific oral presentation delivering. I am grateful you reviewed my manuscripts line by line and feed-backed professional comments. Thank you for pushing me tirelessly to work hard and helping me through these fulfilling two years. I am also grateful for the philosophy of life you told me beyond research when I was in frustrations.

I am thankful to have Dr. James Lacefield and Dr. Ali Khan as members of my advisory committee. Thank you for engaging in my research and for guiding me in the right direction. Your constructive criticisms and words of encouragement have been significant in the development of my research and towards accomplishing my goals.

The past and present members of the Digital Imaging Group of London lab have been immensely supportive providing an invaluable environment to adaptive the abroad life, acquire a wealth of knowledge, and develop as a researcher. To Clara Tam, thank you for being a great friend and colleague. Thank you for helping me to adapt to the new life and study when I first time left my motherland and came to Canada to study alone. Thank you for being there as a teacher to give me advice and guidance patiently, and for imparting your experience to me. To Dr. Chenchu Xu, thank you for being a sincere friend and roommate. Thank you for spending the extra time to aid me in my research. Thank you for putting up with my irritability when I encountered setbacks and teaching me to be pliable but strong. To Dr. Rongchang Zhao, thank you for providing comments to the logical sequence and structure of my papers. To Dr.Liyan Lin, Dr. Rongjun Ge, Dr. Shumao Pang, Mr. Yuqi Qian, thank you for sharing technical knowledge and writing experience.

Most importantly, I would like to thank my parents Zhilu and Xurui. Your support and encouragement have been essential to my success. Although thousands of miles exist between us, you were always my backing to inspire me to do my best, and to have the courage to tackle anything that comes my way.

The computing resources provided through Niagara, funded by the Ontario Government and the Federal Economic Development Agency for Southern Ontario, and local resources provided by Dr. Jaron Chong have been integral to the completion of my research. I would like to express my sincere gratitude and appreciation to them for making my research possible. Computations using the Niagara platform were performed using the data analytics Cloud at the

SciNet (https://www.scinethpc.ca).

Lastly, I would like to express my deepest gratitude to the various sources of funding that I received throughout my graduate studies. I acknowledge funding support from the China Scholarship Council, School of Biomedical Engineering at The University of Western Ontario, and The University of Western Ontario.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-Dimensional
AC	Actor-Critic
AMRL	Anatomy-Modeling Reinforcement Learning
ASER	Adaptive-Sampling Experience Replay
AUC	Area Under the Curve
BP	Back Propagation
CNN	Convolutional Neural Network
CORF	Consequence-Oriented Reward Function
DL	Deep Learning
Dice	Dice coefficient
DRL	Deep Reinforcement Learning
DDRL	Dual-strategy DRL
FC-ResNet	Fully Connected Residual Neural Network
HD	Hausdorff Distance
IA	Image Accuracy
IDR	Identification Rate
IoU	Intersection over Union
KAP	Cohen Kappa Coefficient
MA	Mean Accuracy
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
Loc-Err	Localization-Error
ResNet	Residual Network
PA	Pixel-level Accuracy
ResNet	Residual Neural Network
RF	Radio Frequency
ROC	Receiver Operating Characteristic
RoI	Region of Interest
SCRL	Sequential Conditional Reinforcement Learning
SAC	Soft Actor-Critic
T	Tesla
TCH-ST	Teacher-Student framework
TE	Time to Echo
TR	Repetition Time
USSE	Uncertainty-Sifting Self-Ensembling
VB	Vertebral Body
WSTS	Weakly-Supervised Teacher-Student network
Y-Net	Y-shaped Network

CHAPTER 1

This chapter provides a general introduction to medical object detection and segmentation, as well as concepts about deep learning and deep reinforcement learning. A literature review of current development and challenges of deep learning in medical object detection and segmentation are included in this chapter. The motivation and objectives of applying deep reinforcement learning algorithms in medical object detection and segmentation will be presented.

1 INTRODUCTION

1.1 Overview

Medical object detection and segmentation via Artificial Intelligence (AI) methods are vital roles in computer-aid diagnosis (CAD) to assist radiological experts in clinical diagnosis and treatment planning [1]. Image object detection and segmentation extract the region of interest (ROI), divide a medical image into different areas in pixel-level based on a specified requirement, such as segmenting body organs/tissues in the medical applications for border detection, tumor detection/segmentation, and mass detection[2]. Nowadays, the vast investment and development of medical imaging modalities such as microscopy, X-ray, ultrasound imaging, magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) generate a large amount of medical image data, which attracts researchers to explore AI methods such as Deep Learning (DL) to develop automatic medical object detection and segmentation methods. Applying AI methods to address medical object detection and segmentation has several advantages in CAD: 1) Improving diagnosis efficiency. With the support of high-performance computers, the speed of AI methods detecting and segmenting the medical object from medical images is dozens or even hundreds of times faster than manual labeling by radiologists. AI methods thus speed up the diagnosis process and improve diagnosis efficiency. 2) Increasing the diagnosis accuracy. AI methods process and segment the medical image according to the image content objectively, which avoids inconsistent results caused by observer subjectivity in manual segmentation. AI methods thus increase the diagnosis accuracy compared with manual methods. 3) Reducing the diagnosis cost. AI methods process everything with computers and output the segmentation result automatically. AI methods get rid of the need for experienced radiologists to some extent, which reduces the diagnosis cost for some developing countries with poor medical conditions.

1.2 Medical object detection and segmentation from images

1.2.1 Medical image acquirement

The main type of medical images focused on in this thesis is MRI, which is referred to as cross-sectional slice images as it corresponds to what an object would express if it was sliced open along a special plane. MRI images thus can display detailed information on the anatomy inside the human body.

MRI images are generated by the exchange of radio frequency (RF) energy between the imaging system and a patient's body. This is possible because the most abundant hydrogen atom (protons) within human tissue has unique magnetic properties. More specifically, a proton possesses an innate spin angular momentum rotating about its axis at a constant rate [3–7]. When a strong static external magnetic field is applied, the protons' rotational axes will change from free to align with the magnetic field. Since protons possess an angular momentum, they will precess about the applied field or rotate perpendicularly [8].

The proton's precession rate is proportional to the strength of the magnetic field (see Fig. 1.1), which is expressed by the Larmor (resonance) frequency equation:

$$\omega_o = \frac{\gamma B_o}{2\pi} \tag{1.1}$$

Where ω_o is the Larmor frequency expressed in megahertz (MHz), B_o is the main magnetic field (in Tesla), and γ is the gyro-magnetic ratio expressed in radian per second per tesla (rad·s^{-1}T^{-1}) or MHz/T for $\frac{\gamma}{2\pi}$ [3, 4].

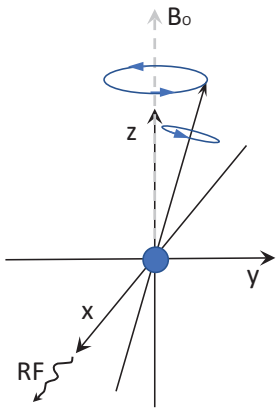


Figure 1.1. Precession of protons in a static magnetic field.

When the MRI imaging acquisition procedure starts, the operator introduces an RF pulse to place the protons into a strong magnetic field, which aligns spinning protons with long axis (parallel/antiparallel).

When the protons' precessing frequency matches the applied RF pulse's frequency, the protons will begin to resonate and absorb some of the energy from the RF pulse, stimulating them into an excited state, which also has an increase in electromagnetic energy. When the operator removes the RF pulse, the protons in an excited state will realign with the magnetic field and release the absorbed electromagnetic energy along the way. This process is known as relaxation. The transmitted electromagnetic energy is then captured by a receiver in the MRI scanner to reconstruct into the MR images.

Because varying tissues in the internal structure of the body contain varying amounts of water, more specifically, the concentration of protons. After the applied RF pulse is removed, MRI detects the emission of the excess energy from the protons to automatically distinguish the intensities of different tissues in the internal structure of the body.

Therefore, the relaxation time (or rate) is the most important factor to generate contrast in different kinds of tissues in the MRI imaging system. It is worth mentioning that the image quality is determined by the intensity of the RF signal.

There are two types of relaxation times T1 (longitudinal relaxation time) and T2 (transverse relaxation time) in MRI. T1 is the time-constant for the spinning protons to realign with the external magnetic field. That is to say, the time it takes for the longitudinal signal to regain 63% of its magnetization value. T2 is the relaxation time-constant for excited protons to lose phase coherence with each other [4, 5, 9]. That is to say, the time it takes for the transverse magnetization to decay to 37% of its original value [10].

MRI acquires image sequences by varying the RF pulses, based on the T1 and T2 relaxation to enhance the quality of select tissues. The repetition time (TR) is the time between successive pulse sequences applied to the same image slice[4, 5, 9]. The time to echo (TE) is the time between the delivery of an RF pulse and receiving the signals emitted from the patient's body, which are referred to as echoes[4, 5, 9].

MRI sequences generate T1-weighted and T2-weighted scans based on the intrinsic T1 and T2 relaxation properties. T1-weighted sequences have shorter TR and TE times, whereas T2-

weighted have longer TR and TE times[4, 5, 9]. T1-weighted images emphasize T1 relaxation, while T2W emphasise T2 relaxation. Fat has Short T1 and medium T2. So that fat shows up bright in T1-weighted and intermediate in T2-weighted. Fluid has Long T1 and T2, so that fluid shows up dark in T1-weighted and bright on T2-weighted.

1.2.2 Medical object detection and segmentation

Medical object detection and segmentation are complex and critical steps in the field of medical image processing and analysis. Their purposes are to detect the regions with certain special meanings (such as lesions and vertebral bodies) from medical images and to classify pixels in the regions into desirable objects, providing a reliable basis for clinical diagnosis and pathology research to assist doctors to make a more accurate diagnosis [11]. The clinical significance of medical object detection and segmentation is (take radiation therapy as an example) [12]: (1) study anatomical structures; (2) identify regions of interest (i.e., locate tumors, lesions, and other abnormal tissues); (3) measure the volume of desirable objects; (4) Observe the growth of the tumor or the decrease of the tumor volume during treatment, and provide assistance for the planning and treatment before treatment; (5) Calculation of radiation dose. To address the medical object detection and segmentation, several methods have been proposed. Traditional medical object detection and segmentation methods involve thresholding, edge detection, region-based, and active contour model methods.

The basic principle of the thresholding method is to calculate one or more thresholds based on the grayscale characteristics of the image, and then compare the image pixel by pixel with the threshold, and finally divide each pixel into the correct category. The thresholding method is based on an assumption of the grayscale image: the grayscale values between adjacent pixels in the target or background are similar, but the pixels of different targets or backgrounds are different in grayscale, which is reflected on the image histogram is that: different targets and backgrounds correspond to different peaks. The selected threshold should be located in the valley between the two peaks to separate the peaks. The disadvantage of the thresholding method is that it is not suitable for multi-channel images and images with similar feature values. It is difficult to obtain accurate object detection and segmentation result where there is no obvious gray difference in the image or there is a large overlap in the gray value range of each object. In addition, because it only considers the gray information of the image without considering the spatial information of the image, threshold segmentation is very sensitive to noise and gray unevenness. Several methods of threshold selection are histogram thresholding method [13], iterative thresholding method [14], and adaptive thresholding method [15]. The most famous

one is the Otsu threshold method (OTSU) [15], which obtains segmentation results by maximizing the variance between classes.

The basic principle of edge detection methods is to achieve segmentation by detecting edges containing different regions. One of the main assumptions of this method is that the change in pixel gray value on the edge between different regions is usually relatively strong. Therefore, according to this assumption, the zero feature of the second derivative of the image and the maximum value of the first derivative can be used to determine whether it belongs to the pixel on the edge. The edge detection method has the advantages of accurate positioning and fast speed, but it cannot guarantee the continuity and closure of the edge and is quite sensitive to noise. Common edge detection algorithms are Canny [16] [17], Sobel [18], and Laplacian [16].

The region-based method is based on directly searching for image blocks, and includes region growth and region separation and merging technologies [19]. The regional growth technique first determines a seed point for each region to be divided as the starting point for growth, and then merges the pixel points in the neighborhood of the seed point that are the same or similar to the point into the area where the seed point is located, and updates the seed point to repeat The above process until no new pixels that meet the conditions are merged. The whole process starts from the seed point, and finally gets the entire area, and then finishes the detection and segmentation task. Splitting and merging can be regarded as the inverse process of region growth. It starts from the whole and continuously splits to obtain each sub-region, and then gradually merges the target regions to achieve the purpose of detection and segmentation. The region-based method is suitable for smooth and uniform images, and the algorithm is simple and the calculation is fast. However, it requires manual selection of seed points, and has poor anti-noise ability. For high-noise, non-uniform medical images, the detection and segmentation results are often poor.

The active contour model uses curve fitting to achieve the purpose of detection and segmentation. First, the contour information is initialized, and then the image information is used to estimate the energy model. The energy is minimized to guide the contour change and gradually approach the target edge to obtain the segmentation result. Snake model [20] is a typical active contour model method, which makes the curve gradually close to the edge under the joint action of image information and the curve itself. It has the advantages of direct interaction with the model, compact model expression and fast implementation speed, but it is very sensitive to the curve initialization and it is difficult to deal with changes in the model topology.

However, with the rapid development of AI, leveraging the artificial neural networks to detect and segment medical objects from images is becoming more and more popular. Compared with traditional methods, AI methods achieve higher precision, faster speed, and higher reliability.

1.3 Deep learning (DL)

Deep Learning (DL) is a subset of machine learning in AI, which imitates the human brain working process to learn to build a non-linear mapping between inputs and the desired output. The learning (or training) procedure of DL is an iterative process (as shown in Fig 1.2): a **convolution neural network** (CNN) predicts an output according to the input(forward propagation); then a **loss function** evaluates the error between the predicted output and the desired output; DL adjusts the CNN's parameters through **back propagation** based on the error to enable its prediction to approach the desired output in next forward propagation. Repeating the above training process, after the error between the predicted output and the desired output is less than a threshold, DL is able to predict the output for a new same category input.

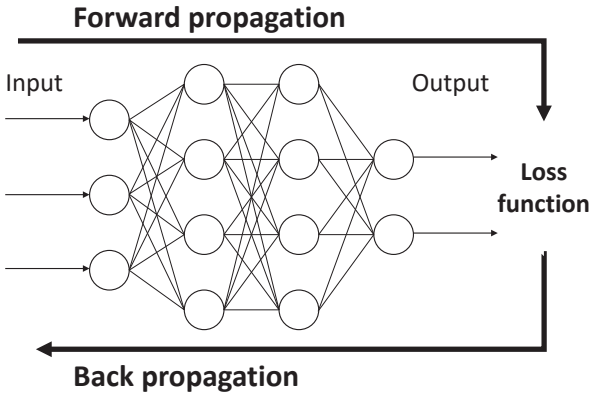


Figure 1.2. The workflow of deep learning.

1.3.1 Convolutional neural network

Convolutional neural network (CNN) [21] assembles low-level features to build abstract high-level features, thereby finding the distributed feature representation in the data. CNN is originated from the artificial neural network, which models human brain neuron network from the perspective of information processing.

Artifical neural network takes neuron as the basic unit and constructs network architecture via the interconnection between neurons. As shown in Fig 1.3, a neuron takes several input

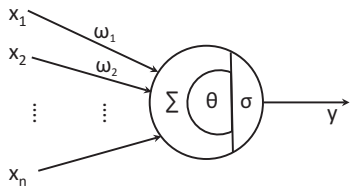


Figure 1.3. An artifical neuron model.

signals x_i and outputs a weighted signal with a bias, which can formulated as follows:

$$y = \sum_{i}^{n} \omega_i * x_i + \epsilon \tag{1.2}$$

Thus a network made up by several neurons can be deployed to building the mapping between an input and an output by learning suitable weights and biases. In imaging processing, convolutions are deployed as neurons to build neural networks, i.e., Convolution Neural Networks (CNNs). Compared with building neurons with pixels, CNN reduces the number of connection weights in neural networks to a range that can be handled in practice. In detail, CNN consists following three layers:

Convolution Layer extracts various image characteristics (such as boundary, texture, and color) through convolution and Stacks convolution sequentially as a hierarchy pyramid to achieve the high-level feature extraction. Convolution is a common mathematical operation in a signal process. A convolution to a matrix is shown in Fig 1.4, the $5 * 5$ square in the left of the figure is regarded as the image input. The $3 * 3$ white square in the middle with the numbers 1, 0, -1, and -2 are the convolution kernels. The convolution kernels have a step size of 1. The sequence moves from the up left corner of the original input to the bottom right corner, and the convolution kernel moves a total of 9-times. The position of the nine time corresponds to the corresponding blue 3 * 3 grid on the right. The number in the grid is the convolution value (here is the result of multiplication and accumulation of the elements in the area covered by the convolution kernel). After 9 movements are calculated, the new matrix of 3 * 3 on the right is the calculation result of this convolution.

In the actual calculation process, the input is an original image and a filter (a set of fixed weights, which is the actual meaning of the convolution kernel mentioned above). After the inner product is obtained, new two-dimensional data is obtained. Different filters will get different output data, such as contour and color depth. If you want to extract different features of the image, you need to use different filters (convolution kernel) to extract the specific information you want about the image. Fig 1.5 shows the process of convolution in a convolutional layer. Through the convolution operation of two different convolution kernels above and be-

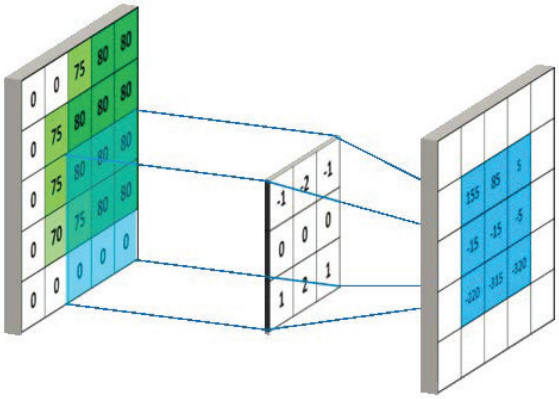


Figure 1.4. A convolution process. The left matrix represents an image before the convolution, the middle matrix represents the convolution kernel, and the right matrix represents the matrix after the convolution.

low, the features of different images can be extracted. The new two-dimensional information will be used as the input of the next convolutional layer in the CNN network. That is, when the next convolutional layer is calculated, the image on the right will be used as the original input image. The shallower (low-level) convolutional layer has a smaller perception domain and learns the characteristics of some local regions; the deeper (high-level) convolutional layer has a larger perception domain and can learn more abstract features. These abstract features are less sensitive to the size, position, and orientation of objects, thereby helping to improve recognition performance.

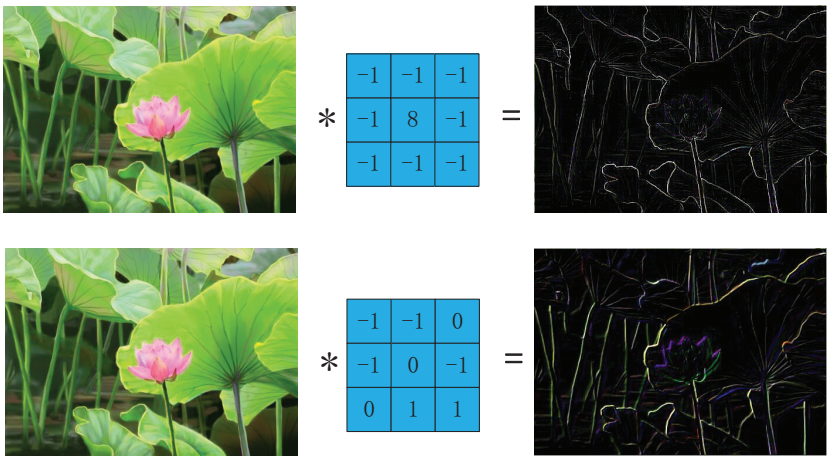


Figure 1.5. Two examples to show different convolution kernels result in different feature maps.

Pooling layer decreases the parameter number of CNN without reducing the image quality.

The feature obtained after convolution is a static feature of the image, which means that a feature useful in one area is likely to be useful in another. At this time, we can use the statistical characteristics of the region to describe the region's information. At the same time, considering that the model needs to have a certain degree of robustness to translation and rotation in practical applications, it is beneficial to enhance the robustness by describing the region features rather than the region itself. There are generally two types of pooling operations, one is average pooling and the other is maximum pooling. The maximum pooling operation is using a 2×2 filter, the maximum is to find the maximum value in each region, here the step size is 2, and finally extract the main features in the original feature map. The method of average pooling layer is to sum each 2×2 area element and divide by 4 to get the main features, Fig. 1.6 illustrates the maximum pooling.

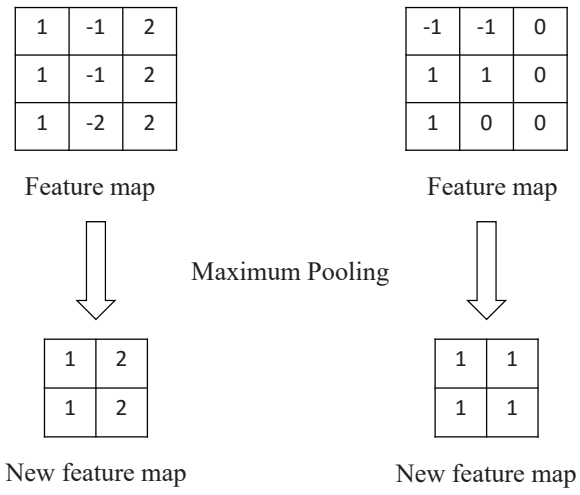


Figure 1.6. An example of pooling process. After the pooling, the new feature map keeps the important information of the original feature map but reduces its size.

Fully Connected Layer is the "classifier" in CNN. If the operations of the convolutional layer and the pooling layer are to map the original data to the hidden layer feature space, the fully connected layer plays the role of mapping the learned feature representation to the sample's label space. In detail, a fully connected layer flats the 2-dimensional feature into a single vector of values, each representing a probability that a certain feature belongs to a label. When multiple fully connected layers are superimposed, a nonlinear relationship between object characteristics and object categories can be constructed.

1.3.2 Loss function & Back propagation

Loss function and back propagation usually work together to enable the CNN prediction to approach the ground-truth. Particularly, loss function measures the distance between the CNN prediction and the ground-truth, back propagation adjusts the CNN parameters based on the loss function value to reduce the distance between the prediction and ground-truth.
Loss function contains following types in this thesis: **MSE loss**, i.e., mean square error, which is usually deployed to measure the segmentation and regression quality. MSE loss calculates the average squared difference between the estimated values and the actual value. If the vector (or matrix) has N prediction $\hat{y}$ and corresponding true value y, MSE loss can be formulated as follow:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1.3}$$

Dice loss which is usually deployed in small object segmentation evaluation. Dice loss calculates the Dice coefficient between the ground-truth and the prediction. Assuming there are a 2D segmentation X and a ground-truth Y, Dice loss is:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1.4}$$

Different from MSE loss, which focuses on the overall performance, Dice loss only pays attention to the segmentation performance on the desirable object. Thus Dice loss is usually employed in small object segmentation to improve the result precision.

Back propagation (BP) is the most common and effective method to train the ANN. The main principle of BP is:
1) Inputting the training data into an ANN and getting the output, which is forward propagation.
2) Because there is an error between the output of ANN and the true label, calculating the error and back-propagating the error to every layer until the input layer.
3) During back-propagation, adjusting the ANN parameters in every layer.
4) Repeating the above procedure iteratively until the error between the output and the true label is less than a tolerance range, which is convergence. The complete formula derivation of BP is shown in Appendix. A.

1.3.3 Types of learning

According to the provided label in a dataset, DL algorithms can be divided into the following categories: fully-supervised learning, semi-supervised learning, and weakly-supervised learn-

ing.

Fully-supervised learning: In the given dataset, the input and label are one-to-one correspondence. Thus fully-supervised learning algorithms aim to train a mapping from the input to the corresponding label. For instance, the label for binary segmentation in the fully-supervised learning is a mask where every pixel is 1 or 0 to indicate the class (foreground or background) of this pixel in the input image. Fully-supervised learning algorithms are the most common training method because theoretically their performance is the best and their training procedures are the simplest.

Unsupervised learning: In the given dataset, there is only input data without any labels. In this case, unsupervised learning algorithms learn the inherent structure of the data from discovering specific patterns through repetitive experiences, such as the gray scale difference between the tumor and the normal tissue. Unsupervised learning algorithms are only suitable for some specific tasks where the data has distinguishable inherent patterns.

Semi-supervised learning: In the given dataset, only partial inputs have corresponding labels, the other partial inputs have unknown labels (or no labels). While semi-supervised learning algorithms aim to make full use of all data in the dataset to learn a mapping from the input the correct label. Generally, semi-supervised learning algorithms are usually used for some tasks where the label is difficult to be annotated.

Weakly-supervised learning: In the dataset, every input has a corresponding weak label. The weak label means it provides less information compared with the label. Weakly-supervised learning algorithms aim to map the input to a more specific label. For instance, in a pixel-level object segmentation tasks, the provided weak label is bounding-boxes of desirable objects. Here, the label is not so accurate that indicates the class of every pixel, but it indicates the object location and size. Thus, the weakly-supervised learning algorithms try to take advantage from these weak labels to exploit the input inherent features, thereby accomplishing the task.

1.4 DL in medical object detection and segmentation

DL achieves great performance in medical object detection and segmentation, but it also exposes many limitations along with the development. Compared with traditional machine learning methods which extract a limited number of image features with hand-crafted algorithms,

DL leverages CNN to extract numerous image features and fully connected layers to regress the object location. More importantly, DL adjusts the parameters of CNN and fully connected layers through BP to extract more proper image features and reduce the distance between its prediction and the ground-truth, thereby achieving great detection and segmentation results. However, DL also has several limitations and cannot address some challenging medical object detection and segmentation tasks.

1.4.1 DL methods for medical object detection and segmentation

DL methods for medical object detection and segmentation can be divided into two categories: 1) Fully convolutional network (FCN) methods; 2) proposal-based methods.

FCN methods are the pixel-to-pixel network in which every layer is a convolutional layer, they input the medical image and output the same size segmentation directly. The most famous FCN method in medical image segmentation is U-Net[22]. Its encoder (downsampling) - decoder (upsampling) structure is a remarkable design method. Particularly, U-Net extracts high-level features from the low-level features through the encoder, and restores the size of the feature to the same size as the input, thereby achieving pixel-level segmentation. Moreover, U-Net also employs skip-connections between the encoder and the decoder to convey the low-level features from the encoder to the decoder directly, which avoids feature losing in the downsampling effectively. There are many new convolutional neural network design methods based on the core idea of U-Net, they add new modules into the encoder-decoder structure or incorporate U-Net with other design concepts, such as Dense U-Net [23], MultiResUNet [24], and Attention UNet [25]. Note that, in such methods, the detection result usually is determined by the surrounding-box of the segmentation.

Proposal-based methods generate one or several Regions-of-Interest (RoIs) as proposals where the desirable object may exist firstly and then segment the RoIs to obtain segmentation. A typical proposal-based method is Mask R-CNN [26], which proposes RoIs on the image, segments each RoI and predicts the bounding-box of desirable objects. Based on the Mask R-CNN, there are lots of proposal-based methods being proposed, such as HMR-Net [27] and Mask scoring R-CNN [28].

1.4.2 Limitation of DL methods

A limitation that DL methods often impose is over-fitting due to their inability to incorporate or discover intrinsic knowledge about the task at hand [29]. In detail, for DL methods, every

aspect related to understanding the task at hand and ensuring the generality of the method are the responsibility of the engineer. While the methods blindly learn to develop a mapping between their input and their label according to the criteria defined by the engineer. As a consequence, such methods often suffer from sub-optimal parameter optimization and weak generalization.

1.5 Deep Reinforcement Learning (DRL)

Deep Reinforcement Learning (DRL) as the newest AI technology has a great potential to address the limitation of traditional DL methods. DRL has a sequential process to interact with the task, which allows DRL to gradually understand the intrinsic knowledge about the task and then complete the task effectively. Therefore, in medical object detection and segmentation, compared with traditional DL methods, DRL can gradually approach the desirable object to locate its location by sequentially interact with the image. This object location facilitates subsequent object detection and segmentation much easier and more accurate.

DRL integrates the advantage of Reinforcement Learning and Deep Learning to achieve the above sequential process. The details about Reinforcement Learning and Deep Reinforcement Learning are introduced as follows.

1.5.1 Reinforcement Learning (RL)

Reinforcement Learning (RL) uses a method similar to trial and error in human thinking to find the optimal strategy. Its self-learning and online learning characteristics make it an important part of the theoretical system of machine learning. In particular, RL deploys an agent to interact with an environment in a proper manner and maximizes long-term rewards. Under an RL situation, the agent is not taught to complete the task, but instead learns to accomplish a task through the reward signal of the feedback. *State*, *action*, and *reward* are the basic element compositions of RL. The state is used to describe the whole of the environment and the agent. The action indicates the role the agent exerts on the environment. At this time, the environment will give the agent a reward signal and a new state. As shown in Fig. 1.7, for the current state s_t, the agent performs an action a_t, gets a reward r_t, and gets a new state s_{t+1}. The core of reinforcement learning is to learn the mapping from state s to action a, that is, what action should be selected in the current state to maximize the reward, which is the policy (strategy). Here, the policy (strategy) can be random, that is, the probability distribution of action to the state, or it can be deterministic. In general, the goal of reinforcement learning is to learn a suitable strategy π to maximize the total cumulative rewards obtained in an episode:

$$R_t = r_t + r_{t+1} + r_{t+2} + \ldots\ldots + r_T = \sum_{i=t}^{T} r_i \tag{1.5}$$

Where T is the iteration number in the episode. Practically, we generally use the sum of rewards with discounts, i.e.:

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots\ldots + \gamma^{T-t} r_T = \sum_{i=t}^{T} \gamma^{i-t} r_i \tag{1.6}$$

Where the discount factor $\gamma \in (0, 1)$ guarantees the sum of rewards can converge when the iteration number is infinite.

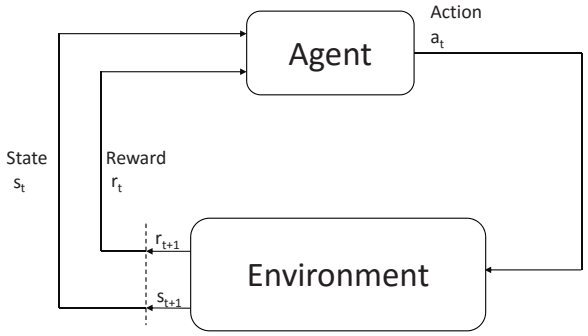


Figure 1.7. The principle of RL.

In RL, a state-value function $V^{\pi}(s)$ is defined to describe the expectation of the sum of rewards that can be obtained by always executing strategy π under state s, namely:

$$V^{\pi}(s) = E\left[\sum_{i=0}^{T} \gamma^i r_{i+1} | s_T = s\right] \tag{1.7}$$

This formula shows the expectation of discounted rewards that can be obtained from the state s, and is also called γ discount cumulative reward.

Similarly, an action-value function $Q^{\pi}(s, a)$ is defined to describe the expectation of the sum of rewards that can be obtained by always executing strategy π after taking action a under state s, namely:

$$Q^{\pi}(s, a) = E\left[\sum_{i=0}^{T} \gamma^i r_{i+1} | s_T = s, a_T = a\right] \tag{1.8}$$

Based on their definitions, the state-value function and the action-value function can be con-

verted to each other by:

$$V^{\pi}(s_0) = \sum_{a \in A} \pi(a|s) * Q^{\pi}(s, a) \tag{1.9}$$

Value-based methods

Value-based methods firstly evaluate the action-value function $Q(s, a)$ for every action, then obtain the optimal policy $\pi(a|s)$ based on the action-value. A typical value-based method is Q-learning. In Q-learning, the action-value function can be obtained by *Bellman Equation* with an iterative solution:

$$Q^{\pi}(s_t, a_t) = E\left[r + \gamma Q^{\pi}(s_{t+1}, a_{t+1})\right]\Big|_{a_{t+1}=\pi(s_{t+1})} \tag{1.10}$$

After knowing the optimal action-value function $Q^{\pi^*}(s_t, a_t)$ for every pair of state-action, the optimal policy (strategy) is to select the action with highest Q value under a specific state, namely:

$$\pi^*(s_t) = \arg\max_a Q^{\pi^*}(s_t, a_t) \tag{1.11}$$

In the beginning of training, $Q^{\pi^*}(s_t, a_t)$ or π^* is unknown, thus the action-value function is trained by iteratively updating the Bellman Equation to get the optimal $Q^{\pi^*}(s_t, a_t)$ or π^*. In addition, there are also many methods estimate the state-value $V(s)$ so that they can obtain the optimal policy (strategy) based on the state-value function

Policy-based methods

Practically, there are infinitely many actions to choose from, it is not feasible to optimize the action-value function (Q function) at this time. To overcome this difficulty, it is more effective to directly learn a parameterized strategy than a value function, Thus, policy-based methods are proposed. A very typical method is the policy gradient method. Policy gradient method uses the derivative $\nabla_w R_t$ of the expected return R_t to the strategy parameters ω as a gradient to directly optimize the strategy parameters. While the derivative $\nabla_w R_t$ can be estimated from samples.

Actor-Critic method

Actor-critic (AC) method combines the advantage of value-based methods and policy-based methods. AC utilizes a parameterized strategy (actor) and a state value function estimator (critic) for improving the strategy. A typical AC algorithm for learning deterministic strategies is the deterministic policy gradient method. The deterministic policy gradient algorithm uses a derivable action-value function approximator, and the policy gradient is obtained by using the derivative $\frac{\partial Q}{\partial a}$ of the output Q value with respect to action a, namely, the policy gradient is $\nabla_{\omega Q} = \nabla_a Q \cdot \nabla_\omega \pi$.

In summary, RL accomplishes a task by interacting with the task environment iteratively (namely, an agent selects actions to change the task environment, and the task environment feedbacks the agent with reward). During the interaction, RL explores the task with various actions, and improves its strategy gradually. Such interaction enables RL to self-learn the optimal strategy and make the optimal decision when dealing with any state of the environment.

1.5.2 DRL algorithms

Deep Reinforcement Learning (DRL) integrates the DL and the RL, thus it combines their advantages has great potential to be applied in medical image analysis. As mentioned above, the performance of RL depends on the process of fitting action-value functions (value-based algorithms) or strategy parameterization (strategy-based algorithms), where the widespread use of function approximators enables RL to be used in more complex problems. While DL has been deployed as function approximators in supervised learning and they are derivable. Thus the DL enlarges the application range of the RL and improves its performance.

In the following part, we introduce two DRL methods: 1) Deep Q-learning Network (DQN) [30], which is a value-based algorithm and has been employed in anatomical landmark detection [29] and breast tumor detection [31]. 2) Soft Actor-Critic (SAC) [32], which is an AC method. SAC is attracting more and more researchers because it is robust to the hyperparameters and performs stably.

Deep Q-learning

Deep Q-learning Network (DQN) is a value-based method, where the input is image information and the output is discrete action (such as the tracking direction in subpixel neural tracking). DQN employs a neural network $Q(s, a|\theta)$ to approximate the optimal action-value function $Q^*(s, a)$ (As shown in Fig. 1.8), where θ represents the network parameter. The optimization

objective of $Q^*(s,a)$ is $y_i = r_i + \gamma Q\left(s_{i+1}, \pi\left(s_{i+1}|\theta'\right)|\theta'\right)$. The network parameter is adjusted by minimizing its loss function $L(\theta)$, namely:

$$\begin{aligned} L(\theta) &= \left(r_i + \gamma Q\left(s_{i+1}, \pi\left(s_{i+1}|\theta'\right)|\theta'\right) - Q\left(s_i, a_i|\theta\right)\right)^2 \\ &= \left(r_i + \gamma \max_a Q\left(s_{i+1}, a|\theta'\right) - Q\left(s_i, a_i|\theta\right)\right)^2 \end{aligned} \tag{1.12}$$

However, since the training target of the action-value function (Q-function) depends on itself, this will cause instability or even divergence in the training process. Here DQN deploys the target network method, that is, make a copy of the original Q-function to get the target Q-function. Then, the original Q-function is trained and continuously updated with the target Q-function as the target, and the target Q-function is copied from the original Q function every certain step. In this way, the complementarity in the training process is removed, thereby greatly improving the stability of the training process.

Another factor that causes training process unstable lies in the strong temporal correlation among the data input to the network, which is because these data (namely (s_i, a_i, r_i, s_{i+1})) is sampled from the same trajectory. DQN solves this problem by experience replay mechanism. As shown in Fig. 1.9, experience replay mechanism stores the exploration experiences into an experience memory, and samples a minibatch of experiences randomly from the memory to train the network. The experience replay mechanism breaks the temporal correlation among input data. It also improves the utilization efficiency of the exploration experiences because every experience is used multi-times.

In addition, a simple η-greedy strategy is employed to ensure the exploration is sufficient. η-greedy strategy means selecting an random action with a probability η. The complete training process can be seen in Appendix. B.

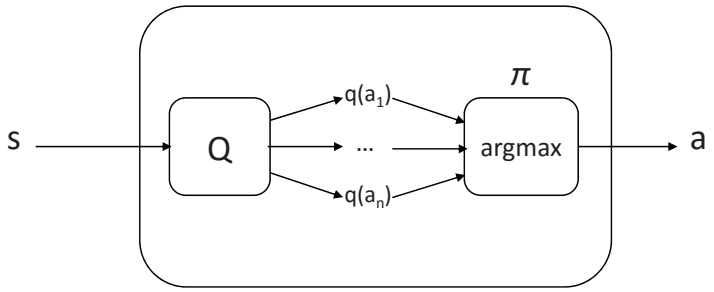


Figure 1.8. The schematic of Deep Q-learning.

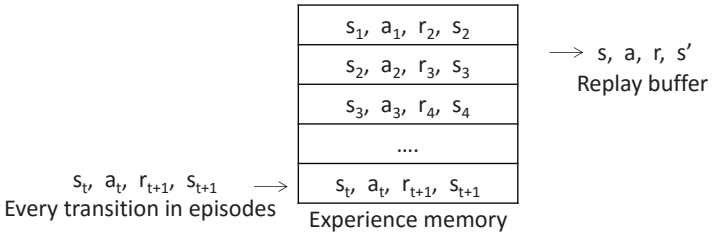


Figure 1.9. Experience replay in the DQN method.

Soft Actor-Critic

Soft Actor-Critic (SAC) [32] is a state-of-art DRL method, it performs much more stable than other methods and is insensitive to the hyper-parameters. SAC employs the framework of Actor-Critic and designs a new optimization objective to train the strategy π:

$$J(\pi) = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \{r(s_t, a_t) + \alpha[\mathcal{H}(\pi(\cdot|s_t))]\} \tag{1.13}$$

Where $\mathbb{E}_{(s_t,a_t)\sim\rho_\pi} r(s_t, a_t)$ is the objective of original DRL, namely, the sum of obtained rewards. Maximizing this item aims to guide the agent to approach and achieve the task goal. $\mathcal{H}(\pi(\cdot|s_t))$ represents the entropy of the strategy, maximizing this entropy aims to guide the agent to explore the liver image as random as possible. In other words, it guides the agent to approach the target goal with different strategies, as well as avoids the agent sticking into sub-optimal.

The training procedure of SAC follows the actor-critic, where a state-value function and an action-value function together play as a critic, a policy (strategy) function plays as an actor. The critic evaluates the actor's strategy according to the feedback of the reward function, the actor improves the strategy according to the critic's evaluation. With the training processing, the critic evaluates the strategy (policy) more accurately, the actor develops a more optimal strategy.

1.5.3 The great potential of DRL in medical object detection and segmentation

DRL is believed to have the ability to understand the medical image through its cognitive-like learning process [29], which can locate the desirable object effectively so that the detection and segmentation can be executed easily. More importantly, its trial-and-error interactions with the task render DRL distinct from traditional DL methods [33]. With a comprehensive exploration to the image, it is promised to capture the intrinsic knowledge and overcome the limitation of

DL methods [34]. More specifically, by combining the strong decision-making ability of RL and the powerful perception of DL, DRL has a sequential process with changing attention-focusing that gradually accumulate evidence of certainty when searching the desirable object [35]. Therefore, in medical object detection and segmentation tasks, with the sequential process, DRL can simultaneously both searching strategy and the appearance of the object of interest to determine the location of the desirable object. With the object location as a prior, the detection and segmentation method can focus on the specific area and obtain a more accurate result. Compared with the traditional DL method which directly predicts the object location with a mapping according to appearance, DRL can avoid falling into sub-optimal effectively. In addition, DRL has shown its competence in other medical image analysis fields [29, 36]. These DRL-based methods have greatly improved the result precision, some of them have become the benchmark in their fields.

1.6 Thesis objective

The overarching topic of this thesis is to leverage the advantage of DRL technology to determine the medical object location and thus facilitate accurate object detection and segmentation. The expectation is to increase the detection and segmentation precision, thereby improving clinical workflow and assisting clinicians to make surgical plans. Therefore, from simple to complex, this thesis deploys DRL into two challenging and representative medical object detection and segmentation tasks: 1) accurate vertebral body segmentation. 2) liver tumor segmentation in the non-contrast-enhanced image. The specific objective for each chapter in this thesis is introduced as follows.

In chapter 2, the objective was to leverage DRL to model the spine anatomy and facilitate vertebral body detection and segmentation from MRI images. The goal was to achieve higher detection and segmentation accuracy than state-of-the-art methods.

In chapter 3, the objective was to leverage DRL to locate the barely-visible liver tumor and facilitate the tumor segmentation in the non-contrast-enhanced image. The goal was to replace traditional contrast-agent-enhanced methods and thus avoid the injection of contrast agents.

In chapter 4, an overview and a summary of the key findings and important conclusion in Chapter 2 and Chapter 3 will be stated. The limitations and new challenges of the methods we proposed will be discussed. The thesis is concluded with potential ideas to generalize DRL to other medical object detection and segmentation tasks.

References

[1] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *Journal of medical physics/Association of Medical Physicists of India*, vol. 35, no. 1, p. 3, 2010.

[2] Y. Guo and A. S. Ashour, "Neutrosophic sets in dermoscopic medical image segmentation," in *Neutrosophic Set in Medical Image Analysis*, pp. 229–243, Elsevier, 2019.

[3] B. M. Dale, M. A. Brown, and R. C. Semelka, *MRI: Basic Principles and Applications*. Wiley, 2015.

[4] A. Berger, "Magnetic resonance imaging," *BMJ (Clinical research ed.)*, vol. 324, no. 7328, p. 35, 2002.

[5] L. Landini, V. Positano, and M. Santarelli, *Advanced Image Processing in Magnetic Resonance Imaging*. Signal Processing and Communications, CRC Press, 2018.

[6] P. Suetens, *Fundamentals of Medical Imaging*. Cambridge medicine, Cambridge University Press, 2009.

[7] S. W. Atlas, *Magnetic Resonance Imaging of the Brain and Spine*. No. v. 1 in LWW medical book collection, Wolters Kluwer Health/Lippincott Williams & Wilkins, 2009.

[8] W. Hei Tse, S. Jin Zhang, and W. Hei, "The design, fabrication, and characterization of nanoparticle-protein interactions for theranostic applications," 2017.

[9] P. Sprawls, *Magnetic Resonance Imaging: Principles, Methods, and Techniques*. Medical Physics Pub., 2000.

[10] G. B. Chavhan, P. S. Babyn, B. Thomas, M. M. Shroff, and E. M. Haacke, "Principles, techniques, and applications of t2*-based mr imaging and its special applications," *Radiographics*, vol. 29, no. 5, pp. 1433–1449, 2009.

[11] D. D. Patil and S. G. Deore, "Medical image segmentation: a review," *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 1, pp. 22–27, 2013.

[12] L. K. Lee, S. C. Liew, and W. J. Thong, "A review of image segmentation methodologies in medical image," in *Advanced computer and communication engineering technology*, pp. 1069–1080, Springer, 2015.

[13] J. M. Prewitt and M. L. Mendelsohn, "The analysis of cell images," *Annals of the New York Academy of Sciences*, vol. 128, no. 3, pp. 1035–1053, 1966.

[14] A. Perez and R. C. Gonzalez, "An iterative thresholding algorithm for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, no. 6, pp. 742–751, 1987.

[15] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[16] P. P. Acharjya, R. Das, and D. Ghoshal, "Study and comparison of different edge detectors for image segmentation," *Global Journal of Computer Science and Technology*, 2012.

[17] M. A. Ansari, D. Kurchaniya, and M. Dixit, "A comprehensive analysis of image edge detection techniques," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 12, no. 11, pp. 1–12, 2017.

[18] O. R. Vincent, O. Folorunso, *et al.*, "A descriptive algorithm for sobel image edge detection," in *Proceedings of Informing Science & IT Education Conference (InSITE)*, vol. 40, pp. 97–107, Informing Science Institute California, 2009.

[19] P. Sharma and J. Suji, "A review on image segmentation with its clustering techniques," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 5, pp. 209–218, 2016.

[20] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] 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.

[23] S. Guan, A. Khan, S. Sikdar, and P. Chitnis, "Fully dense unet for 2d sparse photoacoustic tomography artifact removal," *IEEE journal of biomedical and health informatics*, 2019.

[24] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[25] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[27] C. M. Tam, D. Zhang, B. Chen, T. Peters, and S. Li, "Holistic multitask regression network for multiapplication shape regression segmentation," *Medical Image Analysis*, p. 101783, 2020.

[28] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6409–6418, 2019.

[29] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu, "An artificial agent for anatomical landmark detection in medical images," in *International conference on medical image computing and computer-assisted intervention*, pp. 229–237, Springer, 2016.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[31] G. Maicas, G. Carneiro, A. P. Bradley, J. C. Nascimento, and I. Reid, "Deep reinforcement learning for active breast lesion detection from dce-mri," in *International conference on medical image computing and computer-assisted intervention*, pp. 665–673, Springer, 2017.

[32] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[33] C. Yu, J. Liu, and S. Nemati, "Reinforcement learning in healthcare: A survey," *arXiv preprint arXiv:1908.08796*, 2019.

[34] Y. Man, Y. Huang, J. Feng, X. Li, and F. Wu, "Deep q learning driven ct pancreas segmentation with geometry-aware u-net," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1971–1980, 2019.

[35] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9080–9089, 2018.

[36] G. Luo, S. Dong, K. Wang, D. Zhang, Y. Gao, X. Chen, H. Zhang, and S. Li, "A deep reinforcement learning framework for frame-by-frame plaque tracking on intravascular optical coherence tomography image," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 12–20, Springer, 2019.

CHAPTER 2

The contents of this chapter was previously published in the Medical Imaging Analysis.

2 SEQUENTIAL CONDITIONAL REINFORCEMENT LEARNING FOR SIMULTANEOUS VERTEBRAL BODY DETECTION AND SEGMENTATION BY MODELING THE SPINE ANATOMY

2.1 Introduction

Accurate vertebral body (VB) detection and segmentation from spine images have become a clinically important and frequently demanded research topic for diagnosis and identification of spinal diseases. VB detection usually assists clinicians with the influence caused by VB deformation to the surrounding anatomy, such as evaluating the spinal volume and evaluating the compression of the spinal cord. For instance, VB detection enables clinicians to easily observe the narrowing of the spinal canal caused by VB deformation, thereby assisting clinicians to diagnose the lumbar spinal canal stenosis [1]. VB segmentation usually assists clinicians with VB morphology to diagnose VB-related diseases, such as vertebra collapses/fractures and spondylolisthesis. For instance, a function of VB segmentation is measuring the vertebral body height (VBH) and evaluating the level of vertebral fracture. VB segmentation enables clinicians to easily measure the anterior, central and posterior heights of the observed VB, thereby clinicians can evaluate the difference between these three heights to diagnose the level of vertebral fracture [2]. The same approach, VB segmentation also helps to diagnose the spinal fracture risk for the osteoporotic patients [3, 4] and the diseases related to vertebral morphometry such as crushed/wedged vertebrae [5]. Therefore, VB detection and VB segmentation are helpful to assist clinicians and diagnose spine diseases.

In clinical practice, VB detection and segmentation are completed by radiologists manually delineating a bounding-box and the boundary of VBs, which brings several problems to practical diagnosis: 1) tedious repetition for radiologists caused by manual segmentation for numerous patients. For instance, radiologists manually segment more than 140,000 spine images in the US to diagnose vertebral fractures [6], which takes up a lot of radiologists' time and energy. 2) the inconsistency of the detection or segmentation results caused by the subjectivity of the clinician, especially for an inexperienced clinician. For instance, the clinician's subjective opinion plays a decisive role in the diagnosis of adolescent idiopathic scoliosis (AIS) of 7,000,000 children in the United States [7], which requires correct VB segmentation. 3) prolonged pain caused by delayed diagnoses of vertebral injuries [6]. For instance, 19%-50%

of vertebral fractures of the thoracolumbar spine are associated with a neurological injury [8], which causes great suffering to the patient when the diagnosis period is long.

There are three challenges when obtaining reliable and accurate VB detection and segmentation due to the unique spine imaging and anatomy. **Challenge 1**: mis-detection (or/and mis-segmentation) caused by similar background features from the surrounding tissues. Some of the surrounding tissues in spinal images exhibit similar basic features to the VBs, such as the abdomen region in Fig. 2.1(B) shows similar architecture as the VBs. The adipose tissue is shown with a high pixel-intensity and the belly button possesses a low pixel-intensity, which closely resembles the pixel intensity differences between the VB and intervertebral discs. These similar architectures or appearances are strong enough to mislead the detection and segmentation in medical images. **Challenge 2**: mis-identification caused by the adjacent VBs. Some adjacent VBs are joined together because of some intervertebral disc lesions, such as intervertebral disc degeneration. For instance, the L4 and L5 are jointed so closely in Fig. 2.1(D). It is difficult to define the boundary between these joint VBs relying on the pixel-intensity and prone to treat them as a whole when detecting and segmenting VBs, thereby mis-identifying them as an individual VB. **Challenge 3**: inaccurate segmentation(and detection) caused by the complex appearance and various spatial offsets of VBs. Image artifacts and pathological variations lead to complicated VB appearance and spatial offsets, which imposes great difficulty to accurately detect and segment VBs. As shown in Fig. 2.1(F)-(K), these artifacts and pathological variations totally change the pixel-intensity of VBs in medical images, making it complex to build a robust model and extract the representation features for these VBs. In this case, it is prone to produce false-positive pixels in segmentation masks or deviated detection coordinates.

Although numerous methods have been proposed for VB detection and segmentation, they cannot address the above challenges totally. These methods can be attributed into two types: **1. Proposal-based systems** usually cause false-positive detection and segmentation results. For this type of system, the VB detection and segmentation task is typically decoupled into the proposing of candidate points (or candidate regions) and the analysis of these candidates. In the first stage, these systems propose points (or regions) with a sliding window [9], sampling-manually [10], super-pixels [11], Shannon entropy[12], or integrating two of the above methods [13]. In the second stage, most of the methods in these systems focus on each candidate region and determine which regions are VBs with a classifier, such as support vector machine (SVM) [10], convolutional neural network (CNN) [9, 13] and random forest classifier (RFC) [11]. These methods only pay attention to the local information, namely the information in candidate regions, even if these regions are in the background. Thus these methods lead in many

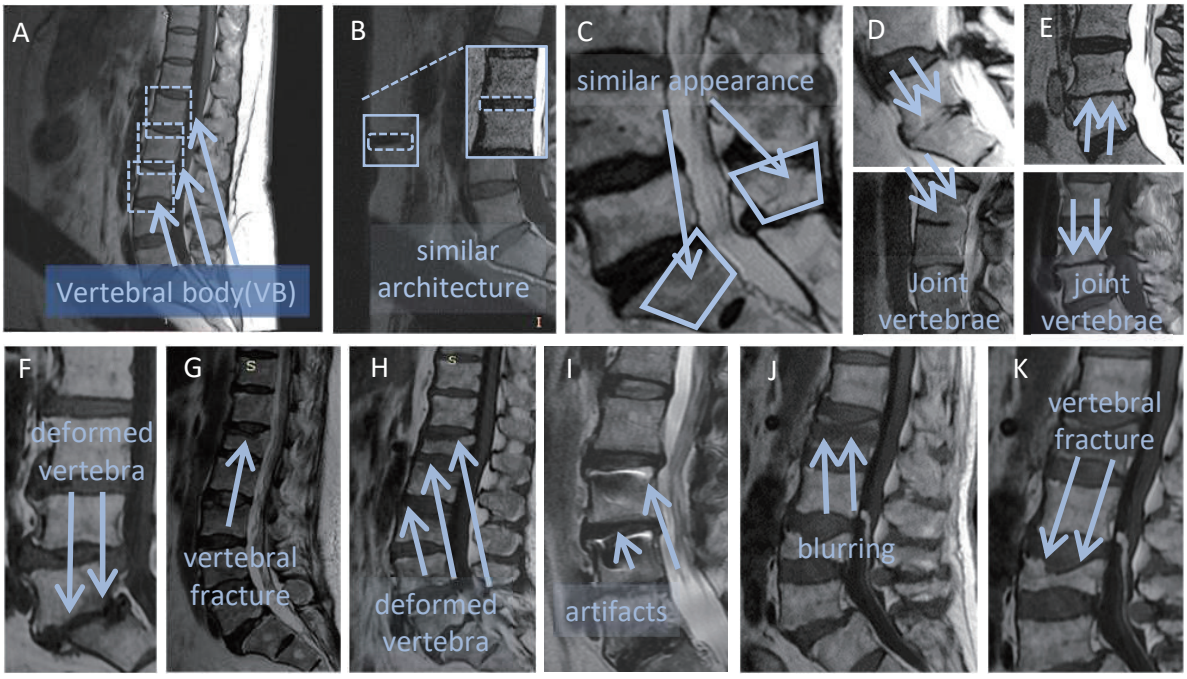


Figure 2.1. Challenges in the automatic vertebral body segmentation and detection. Our method is robust to all these challenges. (A) Vertebral Body(VB); (B)(C) Some background tissues present similar architecture or appearance as VBs (or spines), which misleads the detection and segmentation; (D)(E) Some adjacent VBs joints together because of some intervertebral disc lesions, which leads to these joint VBs being detected and segmented as an individual VB; (F)-(K) Complex appearance and various spatial offsets of VBs caused by image artifacts and pathological variations, which imposes great difficulty to accurately detect and segment these VBs.

cases to false-positive detection and segmentation results. While other methods in candidate-level systems, such as [12], uses a graph model to filter out the background and keeps true candidates as VB center points. Such methods do not consider the influence of pathological variations when generating candidate points (or candidate regions). Therefore, the detection and segmentation results are deviated or false-positive. **2. FCN-based systems** usually cause mis-identification or mis-segmentation. FCN-based systems utilize fully convolutional networks to extract features from entire spine images and to predict every pixel that belongs to the VBs or background, such as Feed-Forward Chain (FFC) proposed by [14], Spine-GAN proposed by [15], and Progressive U-Net proposed by [16]. These methods learn the global features of VBs, such as the S-shape presented by the vertebral column in spine images from the sagittal view, and focuses attention to the vertebral column. As a result, these methods effectively avoid false analysis to the background tissues. However, because these methods lose local-detailed information in the feature-extracting period, these methods cannot perceive the

complex appearance of VBs caused by pathological variations or anatomic variants. Therefore, it is difficult for these methods to separate the attached (joint) VBs [17].

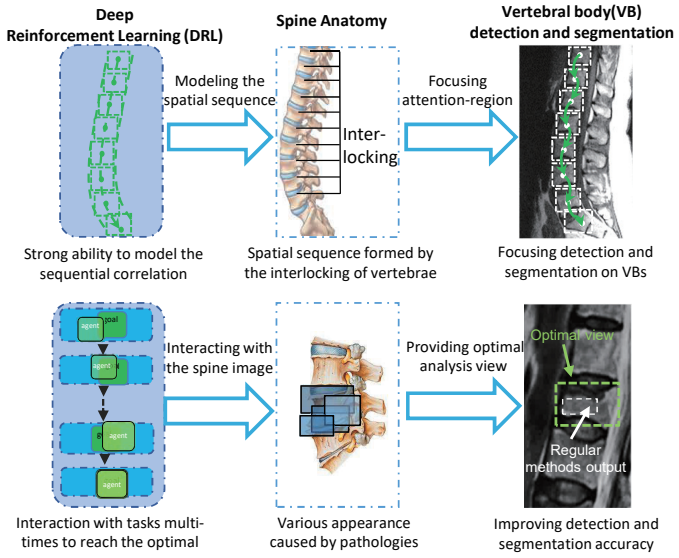


Figure 2.2. Deep reinforcement learning(DRL) has great potential to solve the challenges in VB detection and segmentation: 1) DRL is capable of modeling the spatial sequence correlation of VBs, which proposes an attention-region for each VB and facilitates subsequent detection and segmentation to concentrate on the desirable VB; 2) DRL can explore the surrounding of VBs iteratively and ultimately provide optimal analysis-view for subsequent detection and segmentation. This attention-region settles the VB appearance variation, enables subsequent detection and segmentation networks to achieve unbiased results.

2.1.1 Deep reinforcement learning

Deep reinforcement learning (DRL) [18] has great potential to solve the aforementioned challenges (Fig. 2.2) to achieve accurate detection and segmentation results. 1) DRL has shown great performance on intelligent search [19] and modeling the sequence correlation [20, 21] with its strong decision-making ability, which highly inspires us to exploit DRL to model the spatial sequence correlation of the spine. With modeling the anatomical correlation between the VBs, DRL excludes the interference of the background, such as artifacts and adipose tissue, thereby focusing the attention of subsequent detection and segmentation on the VBs. 2) DRL is believed to gradually accumulate evidence of certainty during exploring its task and ultimately achieves the optimal [22, 23]. Thus DRL can iteratively explore the spine image and gradually understand the surrounding of VBs, thereby providing an optimal analysis view for the subsequent detection and segmentation. The optimal analysis view enables the detection and segmentation to achieve unbiased results.

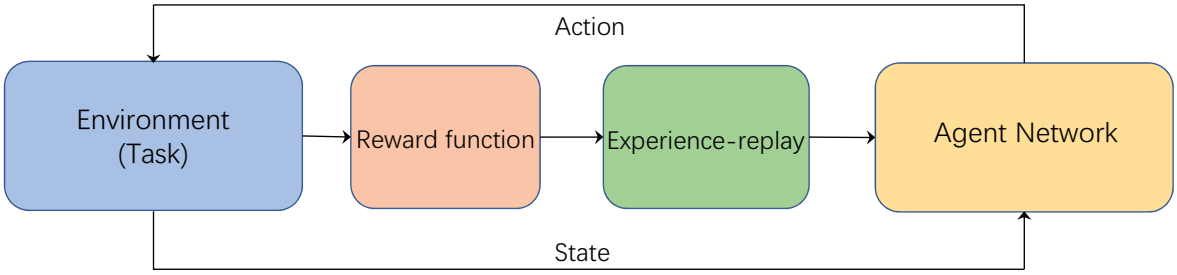


Figure 2.3. The principle of DRL, where the agent observes a state of the environment and makes a decision to select an action from an action set. Each action produces a corresponding reward signal, which acts as a supervised response from the environment. The experience-replay stores these history interaction experiences and replays them to train the agent network, thereby enabling the agent network to select the correct actions to maximize the obtained reward.

Specifically, DRL models its task as a dynamic Markov Decision Process (Fig. 2.3). It consists of: a **state**, an **agent**, an **action**, a **reward function** and an **experience-replay**. The state is an observation of the task. The agent iteratively observes the state of the task and selects a specific action to change the state of the task. The reward function evaluates the action and feedbacks a reward, which guide the agent to select the correct actions and complete the task. The experience-replay is a memory, which stores the these states, actions and rewards into an experience pool, thereby sampling experiences and training the agent to select correct actions to obtain the reward. Therefore, with the supervision of rewards, after iteratively dynamic interaction, the agent can understand the task and select actions to achieve the target.

However, existing DRL methods have three issues when applying the detection and segmentation of VBs: **Issue 1. Unbalanced-experiences.** In spine images, the number of pixels in the background is much larger than the pixels in VBs. Thus, most of the exploration is located in the background(negative-experiences) rather than in the VBs(positive-experiences). This kind of unbalanced-experiences leads to it is time-consuming to model the spatial sequence correlation of spine and search the optimal analysis view. **Issue 2. Reward-confusion.** Because the spine dataset is small compared with other datasets that can be explored infinitely, such as video game data and driving video data. Therefore, with the DRL model training, the sum of obtained rewards becomes complicated, which often causes DRL to struggle in local optimal and training oscillation. **Issue 3. Fixed-exploration.** Because the appearance of VBs is various and complex, it is difficult to apply a fixed exploration mode to handle all VBs in the dynamic decision process. The fixed exploration mode is caused by the action step-size to be

preset and fixed in current DRL methods. Therefore, this makes it difficult or impossible for DRL model to obtain the optimal analysis view for all VBs.

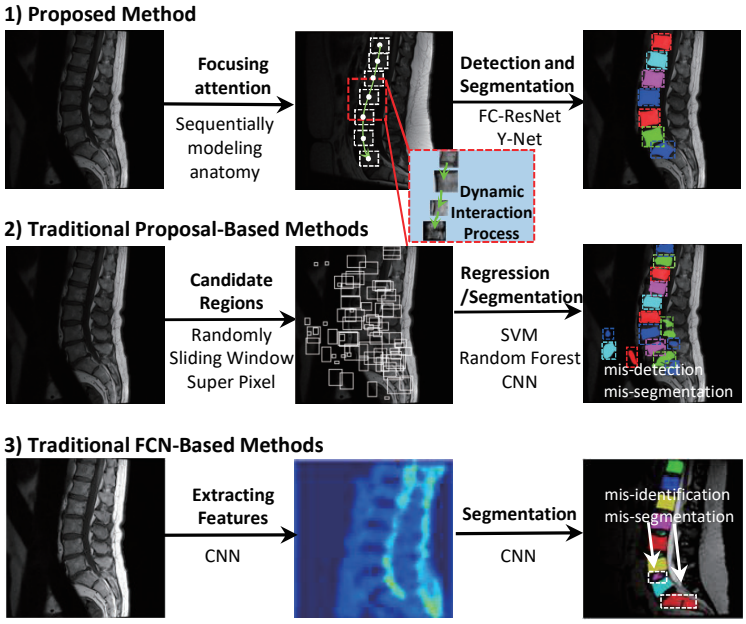


Figure 2.4. Different from traditional methods, our method models the sequential anatomical correlation of the spine and proposes attention-regions for the VBs sequentially, then detects and segments the VBs accurately. SCRL draws support from the spatial anatomical relation of the VBs, SCRL excludes the interference from background tissues on VB detection and segmentation. Simultaneously, SCRL also perceives the geometry and appearance of each individual VB precisely to handle the unpredictable pathological variations or anatomic variants.

2.1.2 Method overview

In this paper, we propose a Sequential Conditional Reinforcement Learning network (SCRL) to tackle the problems of simultaneous detection and segmentation of VBs from sagittal view spine images. SCRL innovatively models the detection and segmentation of VBs from top to bottom in spine images as sequential dynamic interaction processes (Fig. 2.4). Particularly, SCRL firstly proposes an Anatomy-Modeling Reinforcement Learning Network with a novel reward function and an adaptive experience-replay mechanism. This network dynamically models the spatial correlation between adjacent VBs and focuses an attention-region on each VB. Then a new FC-ResNet learns rich global context information of the VB including both the detailed low-level features and the abstracted high-level features to detect the accurate bounding-box of the VB based on the above attention-region. Simultaneously, a new Y-Net learns comprehensive detailed texture information of VB including multi-scale and coarse-to-fine features to segment the boundary of VB from the attention-region. To address the three is-

sues above, SCRL solves the *Unbalanced-experiences* by a newly designed Adaptive-Sampling Experience Replay (ASER), which automatically samples and increases the difficulty of training data. SCRL solves the *Reward-confusion* by a newly proposed Consequence-Oriented Reward Function (CORF), which stimulates DRL model to generate more positive experiences by exponentially connecting the reward of DRL and the attention-focusing accuracy. SCRL solves the *Fixed-exploration* by leveraging Soft Actor-Critic (SAC, Sec. 2.2.1) [24], where the step-size of the action is adaptively decided by the practical distance between the attention-region and the desirable VB.

2.1.3 Contributions

Our major contributions are summarized as follows:

- To the best of our knowledge, our method is the first that leverages DRL to detect and segment VBs in spine images. This DRL-based method effectively models the spine anatomical correlation, thereby achieving better performance than traditional methods.
- Our proposed Consequence-Oriented Reward Function (CORF) addresses the *Reward-confusion* and stimulates DRL to propose an accurate attention-region on each VB. This attention-region enables subsequent detection and segmentation to concentrate on the desirable VB and achieve higher precision. CORF achieves the above goal by for the first time building an exponential relation between the attention-region accuracy and the DRL learning objective, which is inspiring for future work to exploit the continuous action-space DRL method.
- Our new Adaptive-Sampling Experience-Replay (ASER) addresses the *Unbalanced-experiences*, which is a long-standing problem that hinders DRL from applying to small data tasks. ASER effectively promotes the interaction efficiency between DRL and tasks, greatly boosting the convergence speed of the DRL model and increases the attention-region accuracy. ASER achieves the above goals by adaptively increasing the difficulty of the training-data according to the decision-making ability of the DRL model. It is worth mentioning that, ASER can be easily applied to other frameworks where DRL is applied to small medical image datasets.

2.2 Background review

2.2.1 Deep Reinforcement Learning

Deep reinforcement learning (DRL) integrates the powerful understanding ability of deep learning (DL) in perception issues, as well as the powerful ability of reinforcement learning

(RL) in decision-making. DL is one of the most important part of machine learning, by significantly promoting the progress of medical image parsing [25–27]. DL mainly uses the deep convolutional neural network (CNN) as an image feature extractor [28] and a universal nonlinear function approximator [29]. The architecture of CNN is inspired by the feed-forward type of information processing in the early visual cortex of animals [30], it exploits local spatial correlations of image voxels to capture discriminative image features [19].

Inspired by behavioral psychology, RL is a computational method that an artificial agent learns to maximize the cumulative reward signal to reach predefined goals by interacting with an environment [31]. The agent observes a state in the environment and makes a decision to select an action from an action set. Each action produces a corresponding reward signal, which acts as a supervised response from the environment. This reward-based decision making can be formulated as a Markov Decision Process (MDP, [31] $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$), where: $\mathcal{S}$ represents a finite set of states; $\mathcal{A}$ represents a finite set of actions that allow the agent to interact with the environment; $\mathcal{T}$ represents the state transition function, where $\mathcal{T}_{s_t,a}^{s_{t+1}}$ denotes the next state $s_{t+1} \in \mathcal{S}$ is performing the action $a_t \in \mathcal{A}$ in the current state $s_t \in \mathcal{S}$. The objective of standard reinforcement learning is to derive (or approximate) an optimal policy $\pi(a_t|s_t)$ to maximize the obtained amount of rewards $\sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t, a_t)]$, where $\rho_\pi(s_t, a_t)$ denotes the state-action marginal of the trajectory distribution induced by the policy $\pi(a_t|s_t)$. Generally, DRL defines two functions to obtain the optimal policy: 1) the state value function $\mathbf{V}^\pi(s)$ represents the obtained amount of rewards of the policy π that starts from s. 2) the state-action value function $\mathbf{Q}^\pi(s, a)$ represents the obtained amount of rewards of the policy π that starts from s and performs action a. Therefore, the optimization objective of RL can be formulated as $\pi^* = \text{argmax}_\pi \sum_t \mathbf{V}^\pi(s_t)$ or $\pi' = \text{argmax}_{a\in A}\, \mathbf{Q}^\pi(s, a)$.

DRL combines DL with RL to utilizes DL to approximate the tabular function in RL, such as the state value function and state-action value function. Because the action-space and state-space in real-world implementations are high-dimensional, the tabular function in traditional RL cannot characterize such difficult tasks.

Recently, DRL has been applied in some medical image analysis tasks. For medical image registration, [32] casted the image registration problem as a strategy learning process and [33] modeled it as a Markov decision process (MDP). For finding standardized view planes in 3D image acquisitions, [34] utilized DRL to mimic the navigation processes.

Actor-Critic is a typical continuous action-space DRL method. As its name indicates, Actor-

Critic consists of two networks, namely, an actor network learns the policy to select an action, a critic network evaluates the action and gets the state-action value function for the actor to update the policy. Particularly, the actor network approximates the policy function: $\pi_\theta(s,a) = P(a|s,\theta) \approx \pi(a|s)$. The critic network approximates the state-action value function: $\hat{q}(s,a,\omega) \approx q_\pi(s,a)$. With the state-action value function, the actor network updates its parameters with:

$$\theta = \theta + \alpha \nabla_\theta log \pi_\theta(s,a) Q(s,a,\omega) \tag{2.1}$$

Soft Actor-Critic (SAC) [24] is a state-of-art DRL algorithm. SAC integrates Actor-Critic framework and the maximum entropy reinforcement learning framework, which increases the stability, robustness, and exploring ability. Its optimization objective is formulated as:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_\pi \left[\sum_t r(s_t,a_t) - \alpha log(\pi(a_t|s_t)) \right] \tag{2.2}$$

According to Equation 2.2, the optimal policy not only maximizes the expected sum of reward (first summand) but also the expected entropy of itself (second summand), which encourages the strategy decided by the policy to be as random as possible.

2.3 Method

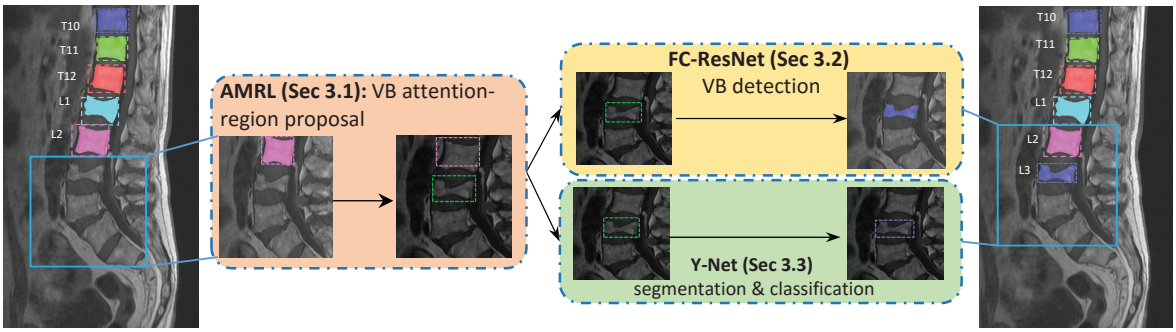


Figure 2.5. The process that SCRL segments and detects L3 vertebral body(VB) based on L2 VB. SCRL coordinates three major components (AMRL, Y-Net, and FC-ResNet) to detect and segment the desirable VB with a coarse-to-fine approach: AMRL proposes an attention-region to the desirable VB, which excludes background interference to subsequent detection and segmentation. Y-Net and FC-ResNet draws support from the attention-region concentrating on perceiving the feature of the desirable VB to detect and segment the VB accurately.

Sequential Conditional Reinforcement Learning (SCRL, Fig. 2.5) integrates three major components to detect and segment VBs simultaneously with a coarse-to-fine approach from top to

bottom along the spine:

Anatomy-Modeling Reinforcement Learning Network (AMRL, Sec. 2.3.1) innovatively adopts DRL to propose an attention-region of the VB. This network accurately obtains an attention-region that includes the VB through Consequence-Oriented Reward Function and Adaptive-Sampling Experience-Replay to respectively effectively and stably model the spatial relationship among the image, the spine, and VBs. The newly proposed Consequence-Oriented Reward Function guides the agent network to explore the image and approach the VB effectively. The newly designed Adaptive-Sampling Experience-Replay stably speeds up the network training efficiency and robustness.

Fully-Connected Residual Neural Network (FC-ResNet, Sec. 2.3.2) learns a transformation matrix, thereby transforming the tight bounding-box coordinate of the desirable VB according to the above attention-region coordinate. FC-ResNet combines the deep abstract-feature extracting of ResNet and the strong function-fitting of fully-connected layers. It perceives the location and size of the desirable VB from low-level to high-level comprehensively, as well as regresses the transformation matrix to obtain the accurate detection result. This accurate detection result is also employed as the start for AMRL when AMRL focuses the attention-region onto the next VB.

Y-shaped network (Y-Net, Sec. 2.3.3) segments the VB mask from the above attention-region, as well as determines the VB classes(lumbar or sacrum). Y-Net learns the texture and geometry information of the VB from multi-scale to obtain the VB segmentation by taking advantage of a U-Net architecture network. By extending the encode-path of U-Net, Y-Net also learns the semantic information from coarse to fine to classify the VB category (lumbar or sacrum).

2.3.1 Anatomy-Modeling Reinforcement Learning (AMRL)

As formulated in Equation. 2.3, AMRL focuses an attention-region on the desirable VB by training an agent network **D** with a MDP-based dynamic interaction process $\{\mathbf{S}, \mathbf{A}, \mathbf{R}, \mathbf{E}\}$. The Multi-Channel State **S** (Sec. 2.3.1) is an observation set obtained from an adaptive attention-region. The Continuous-Transforming Action **A** (Sec. 2.3.1) is the set to adjust the location and size of the attention-region in the spine image freely. The agent network **D** is a DL-based network, its input and output are the state $s \in \mathbf{S}$ and action $a \in \mathbf{A}$ respectively. The Consequence-Oriented Reward Function **R** (Sec. 2.3.1) evaluates that the contribution of the action for focusing the attention-region on the desirable VB and feedbacks a reward r. The Adaptive-Sampling Experience-Replay **E** (Sec. 2.3.1) is a memory, which stores the state s, the action a and the reward r to refer the further agent network training. With a dynamic interaction process (Fig. 2.6), the agent network iteratively observes the state of the attention-region and selects actions to adjust the region location and size, as well as gets reward feedback. The

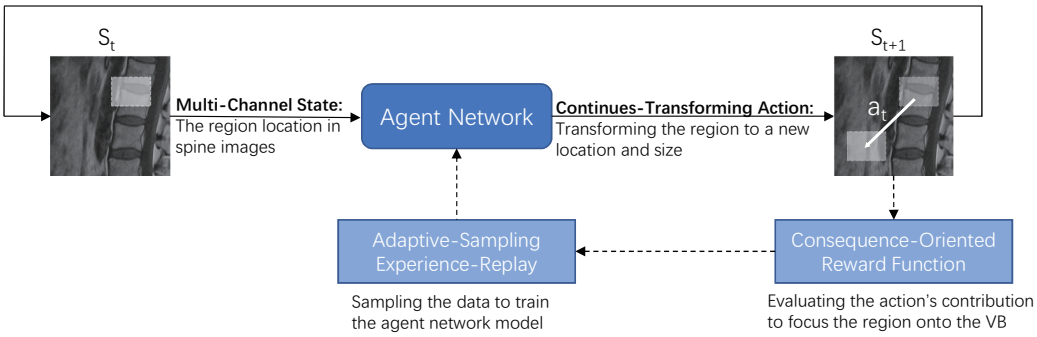


Figure 2.6. AMRL iteratively interacts with the spine image and ultimately focuses the attention-region on the desirable VB. Particularly, the Multi-Channel State of a region as the input is iteratively fed to the agent network. Simultaneously the Continuous-Transforming Action as the output of the agent network transforms the attention-region to a new location and size. Consequence-Oriented Reward Function evaluates the value of the action to the VB attention-region accuracy and feedbacks reward. Adaptive-Sampling Experience-Replay stores the history interaction experiences with the experience pool and samples data to train the agent network.

Adaptive-Sampling Experience-Replay samples training-date and updates the agent network parameters. By combining the above four components dynamically, the agent network selects actions to focus the attention-region on the desirable VB gradually.

$$\begin{cases} a_t = \mathbf{D}(s_t) \\ s_{t+1} \leftarrow s_t \bigoplus a_t \\ r_t = \mathbf{R}(s_t, a_t, s_{t+1}) \\ E \leftarrow E \cup (s_t, a_t, s_{t+1}, r_t) \\ \mathbf{D} \leftarrow Update(\mathbf{D}, E) \end{cases} \tag{2.3}$$

Multi-Channel State

Multi-Channel State (Fig. 2.7) is the input of the agent network in AMRL. It conveys the agent network with the location and size of the attention-region in the spine image in real-time, which enables the agent network to flexibly interact with the spine image through the attention-region. Multi-Channel State is specially designed as three channels. **Channel 1)** a spine image patch, which contains the desirable VB and provides the agent network with comprehensive original information of the VB and the VB surrounding. **Channel 2)** a feature map, which is extracted from the third-last layer in a pre-trained adapted U-Net [35] (pretraining details see Sec. 2.4.3). The feature map contains basic features of the VB and avoids the interference from background

tissues, which improves the agent network learning efficiency. **Channel 3)** a region, which is embedded in a binary image, where the pixel-intensity in the current region is set to 1 and other pixel-intensity outside the region is set to 0. This region plays a role as the attention mechanism, which allows the agent network to focus on a specific area in the spine image. In addition, this region can be transformed to another location and size by the agent, thereby allowing the agent network to focus on different regions in the image. Note that, to make the feature map cooperate tacitly with the third binary region and provide extra VB information, the pretraining of the adapted U-Net takes the binary region into account. By combining the above three channels, our Multi-Channel State facilitates the dynamic interaction between the agent and the spine image, which allows AMRL to understand the spine image comprehensively and search the optimal attention-region.

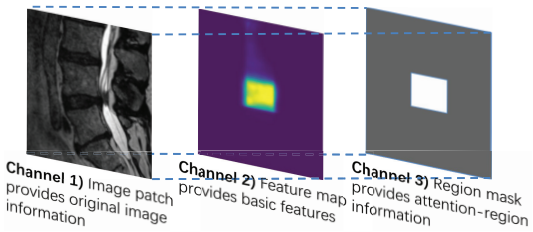


Figure 2.7. Multi-Channel State in AMRL is a concatenation of a spine image patch and its feature map, as well as a region mask. The spine image patch provides the agent network with comprehensive original information of the VB and the VB surrounding. The feature map contains basic features (especially the shape feature) of the desirable VB and avoids interference from the background, which improves the agent network learning efficiency. Particularly, the black purple area in the displayed feature map is background, the green area indicates the attention-region, while the yellow area represents the VB area. The region mask provides the attention-region information, which allows the agent network to focus on the spine image in the attention-region and interact with the spine image.

Giving a spine image, AMRL divides the image into image-patches and initializes the attention-region through following principles: 1) For the image-patch of the first VB, as shown in a) in Fig. 2.8, AMRL takes the top-edge-midpoint of the spine image as a coordinate origin and crops an 196*196 area and a 48*48 area respectively as the image patch and the initial attention-region (selection principles for the above sizes in Sec. 2.4.4). As shown in b) in Fig. 2.8, such image-patches allows the AMRL to transform the attention-region on the first VB, thereby facilitating the further VB detection and segmentation. 2) For other image-patches of other VBs, as shown in c), e) and g) in Fig. 2.8, the AMRL takes the top-edge-midpoint of previous detection VB (the bounding-box) as coordinate origin and crops a 196*196 area as the image-patch. The AMRL takes the bounding-box of previous VB as the initial attention-

region. As shown in d) and h) in Fig. 2.8, such an image-patch and an initial attention-region allow the AMRL to take advantages of the shape similarity between adjacent VBs and focus the attention-region from the previously detected VB on the next desirable VB, thereby facilitating the VB detection and segmentation. Note that, the end signal of sequentially dividing image-patches along the spine is the Y-Net classifies the desirable VB as sacrum because sacrum is the last VB to be segmented and detected. Moreover, to avoid an infinitely sequential dividing to the spine image caused by the error of the Y-Net, we set the maximum number of image-patches is 9, because the maximum number of VBs in a spine image is 9 in our task.

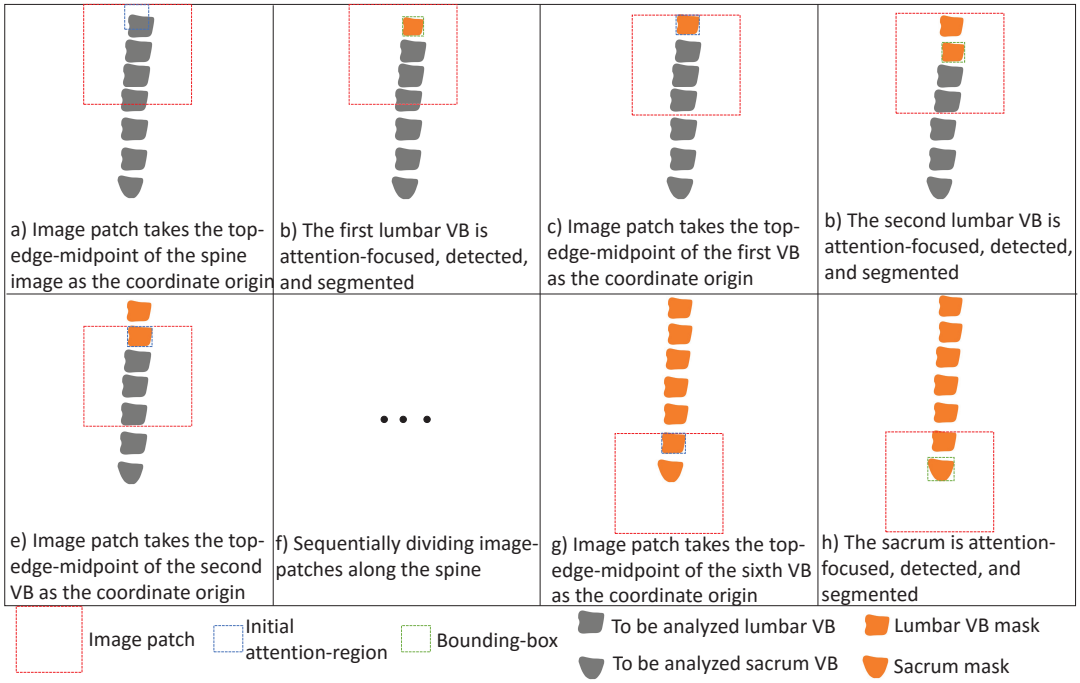


Figure 2.8. Giving a spine image, for the first VB, the image patch takes the top-edge-midpoint of the spine image as a coordinate origin and crops a 196*196 area. For other VBs, the image patch takes the top-edge-midpoint of previous VB (the detection bounding-box) as coordinate origin and crops a 196*196 area.

Continuous-Transforming Action

Continuous-Transforming Action $a \in \mathcal{A}$, as the output of the agent network in AMRL transforms the attention-region to a new location and size(Fig. 2.9), thereby implementing the decision-making of the agent network on the attention-region.

The advantages of our Continuous-Transforming Action is: 1) step-size of actions(both translation and size-changing) is adaptive and variable according to the respective distance between

the attention-region and the desirable VB. This flexibility allows AMRL to balance the step-number and the step-size of actions, and helps the agent network achieve the optimal efficiently. 2) different from traditional DRL methods, our action is multi-dimensional and it changes the location and size of the attention-region simultaneously. This avoids AMRL from redundant action steps when it focuses attention on the desirable VB.

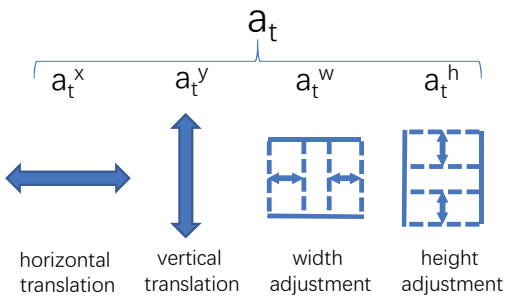


Figure 2.9. Continuous-Transforming Action transforms the attention-region to a new location and size. Continuous-Transforming Action contains four variables, which adjusts the location and size of the region simultaneously. Therefore, the Continuous-Transforming Action implements the decision-making of the agent network.

Continuous-Transforming Action contains four variables: $\mathcal{A} = \left\{a^x, a^y, a^w, a^h \,|\, \mathbb{R}^{a^x,a^y} \in [-6,6], \mathbb{R}^{a^w,a^h} \in [-4,4]\right\}$. $[a^x, a^y]$ translate the attention-region to a new location in horizontal and vertical planes simultaneously, the translation step-size in each plane is $|a^x|$, $|a^y|$ respectively, and the translation direction (left/right and up/down) depends on $sign(a^x)$, $sign(a^y)$ respectively; $\left[a^w, a^h\right]$ adjusts the size of the attention-region, the step-size and direction are similar to the above translations. Therefore, an attention-region $[x, y, w, h]$ is transformed to a new status $[x_n, y_n, w_n, h_n]$ with action $a_t = \left[a_t^x, a_t^y, a_t^w, a_t^h\right]$:

$$\begin{cases} x_n = x + a_t^x \\ y_n = y + a_t^y \\ w_n = w + a_t^w \\ h_n = h + a_t^h \end{cases} \tag{2.4}$$

Agent network

Agent network learns to select actions according to the input state to focus the attention-region onto the desirable VB. The agent network follows the composition modules of Soft Actor-Critic [24], namely it consists of two modules. 1) A policy network approximates the policy $\pi_\theta(s, a)$. This policy outputs a specific action a when observing a state s. Thus the learning

objective of this policy network is increasing the possibility of the actions that can increase the VB attention-focusing accuracy. 2) A soft Q-function network boosts the policy network learning by approximating the state-action value function and updating the network parameter(Equation. 2.1). With the acceleration of the soft Q-function network, the policy network converges to the optimal efficiently. The architecture of the policy network and soft Q-function network are shown in Fig. 2.10.

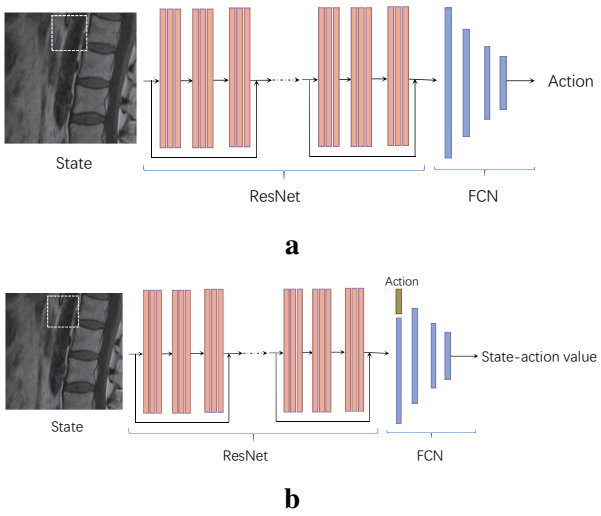


Figure 2.10. (a) The architecture of the policy network, which consists of a ResNet and 4 fully-connected layers. The input and output of policy network are the state and the action respectively. (b) The architecture of the soft Q-function network, which consists of a ResNet and 4 fully-connected layers. The input of the policy network is the state and output is an action. The input of the soft Q-function network are the state and the action obtained from the policy network. The output of the soft Q-function network is the state-action value of the action based on the state.

Consequence-Oriented Reward Function (CORF)

CORF is a specifically designed action evaluation function. It feedbacks the agent network with rewards and guides the network to select correct actions to increase the attention-focusing accuracy. It measures the attention-focusing accuracy difference between the attention-region before and after an action, and creatively adds an exponential weight on those actions that promote the attention-focusing accuracy. CORF is defined as follows:

$$R_a(s, s') = \begin{cases} e^{\beta \times I_{s'}} & \Delta > 0 \\ 0 & \Delta \leq 0 \end{cases} \tag{2.5}$$

where I_s represents the measurement of attention-focusing accuracy when state s, which is the intersection over union between the attention-region under state s and the VB bounding-box from ground-truth. $\Delta = I_{s'} - I_s$ means the improvement of attention-focusing accuracy due to the action $a_{(s,s')}$. β is a weight coefficient to increase the importance of consequence-accuracy in this reward function.

Since the learning objective of AMRL is to learn an optimal policy π to select the optimal action, thereby maximizing the expected amount of rewards:

$$\begin{aligned} \pi &= \operatorname{argmax} \sum \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)] \\ &= \operatorname*{argmax}_{s_t,a_t\sim\rho_\pi} \sum P(s_t)R(s_{t-1},a_{t-1}) \end{aligned} \tag{2.6}$$

where $P(s_t)$ is the possibility that state s appears, which is negatively related to the attention-focusing accuracy i_t when s_t (Proof C). Thus, we assumes that:

$$P(s_t) = e^{-\eta \times i_t},\ i_t \sim s_t \tag{2.7}$$

With above assumption and CORF equation, the right part of equation (6) has such relationship with the attention-focusing accuracy:

$$\begin{aligned} P(s_t)R(s_{t-1},a_{t-1}) &= e^{-\eta\times i_t} \times e^{\beta \times i_t},\ i_t \sim s_t \\ &= e^{(\beta-\eta)\times i_t} \end{aligned} \tag{2.8}$$

In equation (8), β is used to increase the weight of attention-focusing accuracy in the reward function. Therefore, CORF increases the value of β until $\beta > \eta$ can be formulated as follows:

$$P(s_t)R(s_{t-1},a_{t-1}) \propto e^{i_t} \tag{2.9}$$

According to equation (9), CORF enables the learning objective of the agent network to learn an optimal policy π to improve the attention-focusing accuracy:

$$\begin{aligned} \pi &= \operatorname*{argmax}_{s_t,a_t\sim\rho_\pi} \sum P(s_t)R(s_{t-1},a_{t-1}) \\ &= \operatorname*{argmax}_{s_t,i_t\sim\rho_\pi} \sum i_t \end{aligned} \tag{2.10}$$

Therefore, CORF has the ability to stimulate the agent network to select positive actions to

increase the attention-focusing accuracy and address the **Reward-confusion**.

Adaptive-Sampling Experience Replay (ASER)

Adaptive-Sampling Experience Replay (ASER) samples the training-data of AMRL and adaptively increases the training-data difficulty as the decision-making ability of the agent network increases. This not only addresses the **Unbalanced-experiences** but also improves the attention-focusing accuracy. ASER achieves the above objective by two measures: 1) Experience stratified storing. ASER keeps a history experience pool E_r like the regular DRL method, namely stores all kinds of history experiences that the agent goes through (namely, the observed state s, the selected action a, the resulted next state s', the obtained reward r, and the achieved attention-focusing accuracy i) in every iteration into an experience pool. This regular pool mixes recent experiences and breaks the temporal correlations of experiences. In addition, ASER also stores advanced experiences into an advanced experience pool E_a, where whether an experience is advanced depends on whether its attention-focusing accuracy is higher than a threshold γ. This advanced pool stores the experience with lower occurrence probabilities and effectively neutralizes the influence of unbalanced-experiences(Equation. 2.11).

$$\begin{aligned} E_r &\leftarrow E_r \cup [s, a, s', r, i] \\ E_a &\leftarrow E_a \cup [s, a, s', r, i],\ i \geq \gamma \end{aligned} \tag{2.11}$$

2) Experience stratified sampling. ASER samples a mini-batch of training experiences from the advanced pool and the regular pool with respective ratios α and $1-\alpha$ to train the agent network. The experience from the advanced pool is rare and positive for the agent network, thus these experiences improve the decision-making ability of the agent network. The experience from regular pool is redundant and less surprising, thus these experiences stabilize the training of the agent network. In particular, ASER increases the accuracy threshold γ and the sample ratio α gradually during training period to continually enable the agent network to become stronger:

$$\begin{cases} \gamma \leftarrow \gamma + \frac{n}{N} \cdot \theta_1,\ \theta_1 \sim \mathcal{N}(0, 0.05) \\ \alpha \leftarrow \alpha + (1 - \frac{n}{N}) \cdot \theta_2,\ \theta_2 \sim \mathcal{N}(0, 0.6) \end{cases} \tag{2.12}$$

Where n represents current training epoch and N represents the total training epochs. Since the accuracy is corresponding to the decision-making ability of the agent (Proof. C), gradually increasing the accuracy threshold γ can filter the experience that is more valuable for further improving the ability of the agent. Increasing the sample ratio of α increases the amount of rare and positive history experiences, which enables the decision-making ability of the agent

network to become stronger and more robust.

A more detailed ASER implementation follows such procedure: 1) Building two empty experience pool (history experience pool E_r and advanced experience pool E_a) before training; 2) Storing every generated experience $[s, a, s', r]$ in every iteration during training into the pool E_r. If the experience is advanced (namely, the achieved attention-focusing accuracy i in this experience is higher than a threshold γ), storing the advanced experience into the pool E_a. 3) Sampling a buffer of regular experiences $[S_r, A_r, S'_r, R_r]$ randomly from the pool E_r and a buffer of advanced experiences $[S_a, A_a, S'_a, R_a]$ randomly from the pool E_a with ratio $(1 - \alpha)$ and α to construct a minibatch of experiences $[S, A, S', R]$ and train the agent network; 4) Replacing the old experience in the two pools with new generated experience, increasing the attention-focusing accuracy threshold γ and sample ratio α as training process.

2.3.2 Fully-Connected Residual Neural Network (FC-ResNet)

FC-ResNet embeds fully connected layers into ResNet to regress an accurate bounding-box including the VB from the attention-region. For the network input, FC-ResNet concatenates the spine image patch (same as the patch in Sec. 2.3.1) and the attention-region which is proposed by AMRL. This input contains the original comprehensive VB information and the coarse location of the desirable VB. For the network structure, FC-ResNet deploys a ResNet and three fully-connected layers. Particularly, FC-ResNet adopts the architecture of ResNet without affecting the detection result, where the numbers of bottlenecks are $[3, 4, 4, 3]$ and the numbers of convolution kernels are $[64, 128, 128, 32]$ for each of the layers respectively. Such architecture adoption prevents ResNet from over-fitting on our spine image dataset. For the regression objective of fully-connected layers, since the attention-region has achieved a high detection accuracy, FC-ResNet regresses the translation factors (t_x, t_y) and scale factors (t_w, t_h) to predict a bounding-box by linearly transforming the coordinates of the attention-region:

$$\begin{cases} t_x = (G_x - R_x)/G_w \\ t_y = (G_y - R_y)/G_y \\ t_w = log(G_w/R_w) \\ t_h = log(G_h/R_h) \end{cases} \tag{2.13}$$

Where (G_x, G_y, G_w, G_h) represents the bounding-box of the desirable VB from the ground-truth, (R_x, R_y, R_w, R_h) represents the attention-region proposed by AMRL. During testing period, FC-ResNet is able to transform the attention-region to the prediction bounding-box with the above

4 transformation factors. Such transformation-factor regression method bounds the regression range to [0, 1], which avoids gradient explosion and makes the FC-ResNet model training more stable.

2.3.3 Y-Shaped Network (Y-Net)

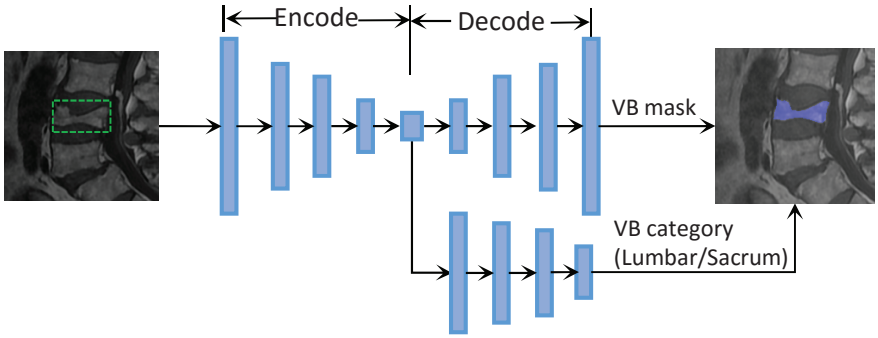


Figure 2.11. The architecture of the Y-Net. Y-Net deploys an encode-decode structure to predict the VB segmentation mask. Simultaneously, Y-Net compresses the encode path further to classify the VB. If the VB is classified as sacrum, the total process to the spine image ends.

Y-Net is an encoder-decoder network with two parallel outputs (segmentation and classification, as shown in Fig. 2.11), which is modified from the U-Net architecture [35]. Y-Net takes advantage of the attention-region proposed by AMRL to exclude the interference from background tissues, and draws support from the strong detail-preserving ability of U-Net to handle the various appearance of VBs. For the network input, Y-Net builds it the same as the FC-ResNet's. For the network architecture, Y-Net deploys an encode-decode architecture to get the VB segmentation. The encode path improves the robustness of the network to disturbance and noise, such as imaging artifacts and blurring. The decode path recovers the abstract features to the original size and outputs the pixel-level segmentation result. The concatenation between encode path and decode path effectively transmits the local features from low-level to high-level, which prevents loss of details and enables Y-Net to handle various VB appearances. In addition, Y-Net appends the encode path as further compression steps and determines the class (lumbar VB or sacrum) of the desirable VB with fully connected layers, which determines the terminal of the sequential segmentation and detection to a spine image because the sacrum is the last VB in the entire process. This further compression to the encode path makes full use of the semantic information extracted by the deep convolution layers since the encoding path is commonly understood as the recognition path[36].

2.4 Data and experiments

2.4.1 Data acquisition

The proposed SCRL has been intensively evaluated with a challenging dataset. The dataset comes from 240 multi-center clinical patients including 80 males and 160 females with an average age of 61 years the oldest being 98 years old, and the youngest being 29 years old. Because these multi-center patients are examined by different vendor models, their MRI scans have different parameters. The range of the repetition time (TR) is from 340 ms to 4,000 ms with mean of 1509 ms; echo time (TE) from 8.072 ms to 147 ms with mean of 70.331 ms; flip angle (FA) is between 90°-180°; slice thickness from 0.88 mm to 4 mm with mean of 2.8823 mm; and in-plane pixel spacing from 0.391 mm to 0.625 mm with a mean of 0.515 mm. For each subject, a midsagittal lumbar MR images are selected manually to conduct VB detection and segmentation. The dataset (121 T1 weight and 119 T2 weight) has 240 S1-L5 vertebrae, 170 T12 vertebrae, and 100 T11 vertebrae. The segmentation and detection ground truth is labeled by our lab tool according to the clinical criterion and double-checked by two experienced spinal radiologists.

It should be noted that although we use the lumbar spine data to train and test SCRL, SCRL is not limited to analyze the lumbar VB, i.e., as long as the data is a sagittal slice of the spine, SCRL can detect and segment all VBs from top to bottom. Because SCRL models the anatomy between adjacent VBs and sequentially approaches each VB, if the spine image contains more VBs, this will promote the effectiveness of SCRL by providing more training data.

2.4.2 Data augmentation for AMRL

SCRL augments training data from the original spine image based on the ground-truth to simulate the practical VB attention-focusing to increase the robustness of AMRL. For the practical attention-region proposing of AMRL, the start attention-region is the bounding-box of the previous detected VB. Therefore, in data augmentation, SCRL adds Gaussian noise $\mathcal{N}(0, 15)$ to the center coordinates and size respectively for the above start attention-region, which makes the start attention-region more complex and improves the decision-making ability of AMRL.

2.4.3 Pretraining U-Net for channel 2 in Multi-channel State

The U-Net is pretrained on image-patches cropped from our spine dataset (testing data not included), where every image-patch attaches an attention-region. Therefore, the pretrained U-Net is a VB feature extractor to focus on the attention-region and provide VB features. In detail,

as shown in the left of Fig. 2.12, the input for the pretrain concatenates an image patch and an attention-region, where the image patch is cropped along the spine and its top-edge-midpoint is the same as every VB's. The attention-region is a box generated randomly in the patch to simulate the random exploration of AMRL, namely, in the practical training of AMRL, this attention-region is the proposal of AMRL. The label for the pretraining is a VB mask-patch cropped from the ground-truth, where the patch has the same size and location as the image-patch and only the mask inside the attention-region is reserved. With such input and label, the pretrained U-Net is highly sensitive to the VB inside the attention-region, which allows AMRL to observe VB features through the attention-region, thereby transforming and focusing the attention-region on the desirable VB.

The third-last layer in U-Net is marked as the red arrow in the right of Fig. 2.12. This layer has the same size as the image-patch and combines multi-level VB information, the former indicates this feature map reserves the spatial information, and the latter indicates this feature map can provide detailed shape information compared with other two channels of our Multi-channel State. Therefore, we decided to extract the feature map from this layer as a part of our state to improve the attention-region-focusing accuracy. In detail, 1) this layer is in the last block in the decode path, it has the same size as the input image-patch (channel one in the state). The same size is the necessary condition that the feature map from this layer can be a part of the state because it avoids resizing and reserves the spatial information. 2) This layer combines low-level VB features from encode-path and high-level VB features from the bottleneck. Thus, the feature map in this layer provides the AMRL agent with more accurate VB information than that provided by the other two channels (image-patch and attention-region).

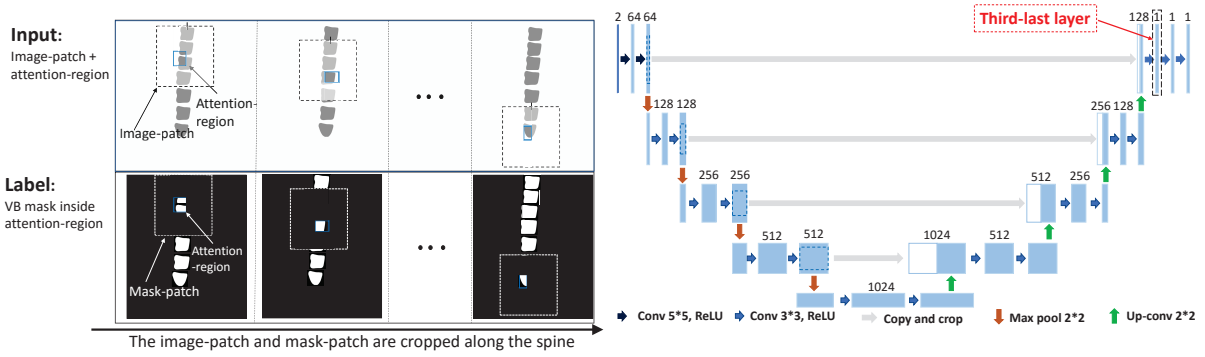


Figure 2.12. The architecture of the adapted U-Net and related training data. The third-last layer is shown as the red arrow. The input is an image patch and a randomly generated attention-region. The supervision is the VB mask inside the attention-region.

2.4.4 Stated-related hyperparameter setting

The state-related size setting (including image patch size 196*196 and initial attention-region size 48*48) are image-dependent. Following are the principle to determine these size values: **1) Image patch size (196*196).** The image patch size decides the state space size and smaller size results in stronger neural network robustness [37]. According to the clinical protocol, the first VB usually locates on the mid-line. However, in some cases, the VB deviates from the mid-line due to lordosis. Thus, as shown in the left of Fig. D.1), the experiment counts the distribution of the first VB and selects 196 pixels as the image patch size, which can cover all the first VB in our dataset. Some extreme examples are illustrated in Fig. D.2. **2) Initial attention-region size (48*48).** The average size (width and height) of the first VB is are 51.63 pixels for width and 45.48 pixels for height (right in Fig. D.1), so that 48*48 is selected as the initial attention-region size, which allows the attention-region to be transformed to the first VB size with few steps and improves the attention-focusing accuracy.

2.4.5 Training strategy for SCRL

The training period of SCRL is divided into two periods: 1)training AMRL, Y-Net, and FC-ResNet respectively. This step enables AMRL the ability to propose an attention-region for the desirable VB, as well as Y-Net and FC-ResNet to be able to segment and detect the desirable VB with the support of an attention-region. 2) Fine-tuning Y-Net and FC-ResNet with the attention-region proposed by AMRL. This step enhances the robustness of Y-Net and FC-ResNet to the attention-region deviation, increasing the compatibility of AMRL, Y-Net, and FC-ResNet.

2.4.6 Implementation details

All the input spine image to SCRL is re-sized to size 512*512 directly without manually cropping. This resizing method causes different pixel spacing (physical size), thus in the late experimental result evaluation, the related pixel spacing also is re-sized. The batch size in AMRL is set to 1 considering the special iteration of AMRL and image normalization. Particularly, to address the various spine image intensity distribution, SCRL uses *Min-Max scaling* to normalize the image, where the min and max intensity values totally depend on this image instead of the whole dataset. The training configurations which are existing in AMRL, Y-Net, and FC-ResNet are listed in Table. 2.1. Other parameter configurations of AMRL mainly include the capacities of regular experience pool and advanced experience pool are 20000 and 5000 respectively; The initial γ and α is 0.6 and 0.0 respectively. The target smoothing coefficient is 0.005, the target update interval is 1. The code of SCRL is implemented with Python 3.6 on

Algorithm 1 Training AMRL

Initialize parameters: α, γ
Build empty experience pools E_a, E_r
Initialize Soft Actor-Critic: soft_q_net, policy_net, value_net
for each epoch **do**
    for each spine image **do**
        image patch initialization
        for each *image patch* **do**
            state $s \leftarrow$ initial_state(*image patch*)
            for each_iteration **do**
                action $a \leftarrow$ policy_net(s)
                reward r,accuracy i, new state $s' \leftarrow$ executing_action(s, a);
                if $i > \gamma$ **then**
                    $E_a \leftarrow E_a \cup (s, a, r, s')$
                end if
                $E_r \leftarrow E_r \cup (s, a, r, s')$
                $S, A, R, S' \leftarrow$ sample a minibatch experiences(E_r, E_a, α)
                update soft_q_net(S, A, R, S')
                update policy_net(S, A, R, S')
                $s \leftarrow s'$
            end for
        end for
        update α and γ
    end for
end for

PyTorch 1.1.0. The parameters of SCRL is initialized using *Glorot initialization* [38] for fully connected layers and *He initialisation* [39] for convolutional layers. The robustness of SCRL is so strong that we tested various hyper-parameters and they all achieved great results. The training of SCRL is implemented on 4 NVIDIA Tesla P100-SXM2 GPUs.

For evaluation and comparison, We divide the data set into a training set, a validation set, and a test set in the ratio of 80%, 10% and 10%. Since the training time of the AMRL is much longer than the training of traditional DL methods, cross-validation is inapplicable in this project. For the training dataset, it is used for training all models in SCRL, including AMRL, Y-Net and FC-ResNet. For the validation dataset, it is used to select the optimal model parameter after training. For the test dataset, it is totally independent of the training process and only used for the test.

Table 2.1. Training configurations of the SCRL network

Network Name	AMRL		Y-Net		FC-ResNet
	soft q-function net	policy net	segmentation	classification	detection
Loss function	MSELoss	-	BCELoss	CrossEntropyLoss	SmoothL1Loss
optimizer	Adam		RMSprop		Adam
Initial learning rate	1e-4		1e-5		
End learning rate	1e-6		1e-6		
Learning rate decay factor	0.95		0.96		
Batch size	120		1		
Number epochs per decay	1				
Drop-out rate	0.2				

2.4.7 Experimental setting

We evaluate SCRL from the aspect of VB detection, VB segmentation and attention-focusing accuracy in the following experiments.

Detection evaluation for FC-ResNet

To illustrate the detection performance of FC-ResNet, three state-of-the-art metrics are computed by comparing the detection result obtained from FC-ResNet detection result with the ground-truth: Intersection-over-Union (IoU), Localization-Error (Loc-Err) [9] and Identification Rate (IDR) [40].

Intersection-over-Union (IoU) is deployed to measure the overlap between the bounding-box by FC-ResNet and VB bounding-box from ground-truth, which is defined as:

$$IoU = \frac{p \cap g}{p \cup g}$$

where p represents the bounding-box by FC-ResNet for the desirable VB and g represents the ground-truth bounding-box of the desirable VB.

Localization-Error (Loc-Err) is deployed to measure the deviation of the bounding-box by FC-ResNet compared with the ground-truth, which is defined as the Euclidean distance between the prediction centroid and the ground-truth centroid:

$$Loc\_Err = \sqrt{(x_g - x_p)^2 + (y_g - y_p)^2}$$

where (x_g, y_g) and (x_p, y_p) are the centroids of ground-truth bounding-box and proposed bounding-

box.

Identification Rate (IDR) is employed to measure the overall detection true positive rate, which is defined as the rate between the number of correctly identified VBs and the total number of VBs:

$$IDR = \frac{\sum_{i=1}^{N} Loc\_Err_i \times Ps_i < \tau}{N}$$

where N is the number of total VBs, τ is a threshold (the original value $\tau = 20$ mm) that determines whether the VB is detected correctly. Ps_i is the pixel spacing of this spine image. By taking the various pixel spacing of spine images into consideration, Loc-Err makes the detection evaluation fair and addresses the influence of pixel spacing difference.

Segmentation evaluation for Y-Net

To evaluate the segmentation performance of Y-Net, three typical metrics are calculated by comparing the segmentation result obtained by Y-Net with the ground-truth: Pixel-level accuracy (PA), Dice coefficient (Dice), and Hausdorff distance (HD).

Pixel-level accuracy (PA) calculates the rate between the amount of right classified pixels and the amount of total pixels, which is defined as:

$$PA = \frac{p \cap g}{g}$$

where p represents the segmentation mask of the desirable VB and g represents the ground-truth mask of the desirable VB.

Dice is a measure of the extent of spatial overlap between the segmentation and the ground-truth, which is defined as

$$Dice = \frac{2TP}{2TP + FP + FN}$$

where TP is the true positives of the desirable VB, while FP and FN are false positives and false negatives of the desirable VB respectively.

Hausdorff distance (HD) is a spatial distance-based metric widely used in the evaluation of

segmentation boundary as a dissimilarity measure, which is defined as:

$$HD(P,G) = max(h(P,G), h(G,P))$$

where P, G denote the boundary point set of segmentation and ground-truth of the desirable VB. While $h(g,r)$ is called the directed Hausdorff distance and is given by $h(g,r) = max_{g\in G} min_{r\in R}\|g-r\|$, where $\|g-r\|$ is Euclidean distance in our evaluation. Note that, here the Euclidean distance is the physical distance instead of pixel distance. By taking the physical distance into consideration, HD makes the segmentation evaluation fair and addresses the influence of pixel spacing difference in difference spine images.

Classification evaluation for Y-Net

To evaluate the classification performance of Y-Net and the overall classification of SCRL, two typical metrics are calculated by comparing the classification result obtained by Y-Net with the ground-truth: mean accuracy and image accuracy.

Mean accuracy (MA) as a common matrix to evaluate classification performance, which is the ratio of number of correct predictions to the total number of input samples:

$$MA = \frac{Numberofcorrectpredictions}{Totalnumberofpredictions}$$

Image accuracy (IA) as a rather strict matrix [9], which measures the ratio of the correctly classified image out of the all image. Here, the correctly classified image indicates an image is considered as correctly classified only if all VBs in this image are correctly classified. IA is defined as following:

$$IA = \frac{Numberofcorrectlyclassifiedimage}{Totalnumberofimages}$$

Attention-focusing evaluation for AMRL

To tune the optimal action-related hyperparameter for AMRL action-related size setting (namely, the number of actions/steps in an episode, the step-sizes of translation and size adjustment), we set different hyperparameter settings according the task demand and compare the attention-focusing accuracy of AMRL. Particularly, the step-sizes of action multiple the number of actions should be larger than the maximum task demand, where the maximum task demand for displacement is 106 (Fig. D.3) and for size-adjustment is 67 (Fig. D.4). Under this condition,

multiple experimental settings are deployed, where the number of actions are set as [15, 20, 25] and the corresponding step-sizes of action are set as 8,6, 6,4, 5,3 respectively. These settings totally meet the task demands.

To show the effectiveness of CORF and ASER in attention-focusing, we cross-combine CORF, ASER, regular reward-function (RR) and regular experience-replay (RER) to train the AMRL and record 1) the mean and minimum attention-focusing accuracy achieved by AMRL with these 4 combinations in the validation dataset; 2) the sum-of-rewards obtained by AMRL when ASER and RER combines CORF respectively in the validation dataset. The action-related hyperparameters are the above tuned hyperparameters. The attention-focusing accuracy is measured by the Intersection-over-Union (IoU, Sec. 2.4.7).

2.4.8 Results and discussion

Detection result

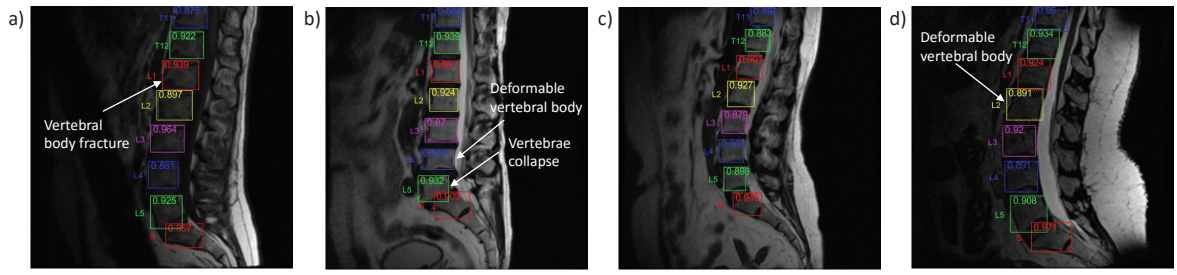


Figure 2.13. Qualitative visualization shows that our SCRL network accurately detects all VBs in the spine images. The green dashed box is the predicted bounding-box and the red dashed box is the ground-truth bounding-box. 1) The IoU between these two kinds of boxes is also provided. There is no mis-detection result on the background, which explains the spatial sequence modeling of AMRL plays an irreplaceable role in VB detection. 2) All the detection result achieves a high IoU, which explains FC-ResNet has the ability to handle various VB appearances. 3) Although there are some deformation caused by the pathology, SCRL still detects them with tight bounding-boxes.

Qualitative evaluation. Fig. 2.13 demonstrates that SCRL achieves excellent detection. The detection results in Fig. 2.13 are of different VB appearance, image resolution and intensity distribution from our validation dataset. The predicted detection boxes (dotted box) have high overlaps with the ground truth boxes (solid box) despite the highly various characteristics of spine images and complicated VB appearance caused by spinal diseases. SCRL is quite effective in the VB detection: 1) There is no mis-detection result on the background, which demonstrates the anatomical correlation modeling of AMRL play an irreplaceable role in VB

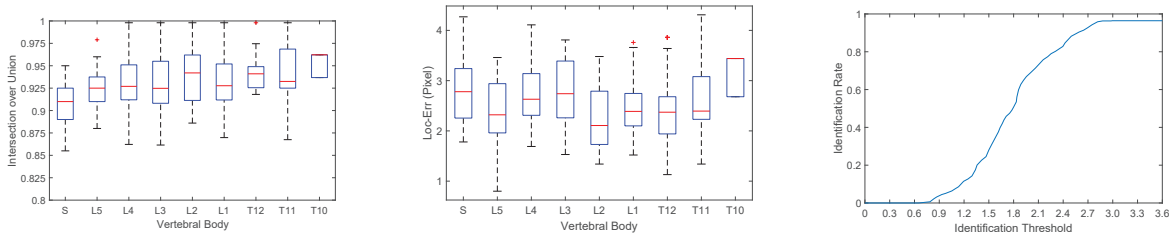
detection; 2) All the detection result achieved a high IoU, which shows that FC-ResNet has the ability to handle various VB appearances caused by pathologies.

Quantitative analysis. Based on the quantitative results in the first row of Table. 2.2, the detection of SCRL on average achieves 92.3% in IoU, 3.142 mm in Loc-Err, and 96.3% in IDR, which indicates the predicted bounding-boxes have high overlaps with the ground truths and the VB centers are predicted accurately.

Table 2.2. Performance comparison of the proposed SCRL with existing state-of-art VB detection methods and VB segmentation method. SCRL significantly outperforms its competitors in all the cases

Methods	Detection			Segmentation			Classification	
	IoU	Loc-Err	IDR	PA	Dice	HD	image accuracy	mean accuracy
SCRL	0.923	3.142	0.963	0.949	0.926	1.943	0.958	0.964
TDCN ([13])	0.818	7.294	0.844	N/A	N/A	N/A	0.833	0.916
Hi-scene ([9])	0.919	5.199	0.963	N/A	N/A	N/A	0.933	0.964
FFC (Single-class, [14])	N/A	N/A	N/A	0.905	0.890	2.362	N/A	N/A
Spine-GAN ([15])	N/A	N/A	N/A	0.922	0.917	2.406	0.917	0.961
SuperPixel-RFC (Single-class, [11])	N/A	N/A	N/A	0.879	0.841	4.260	N/A	N/A

Based on the detection result of each kind VB(Fig. 2.14(a)(b)), SCRL has the ability to handle all kinds of VBs and detect them accurately. For the first VB (such as T10, T11, and T12), SCRL achieves the minimum IoU of 85% and mean IoU of over 92%, the maximum Loc-Err of 5 pixels and mean Loc-Err of less than 3.5 pixels. This great detection result illustrates SCRL is capable of modeling the spatial relation between the first VB and the spine image,



a **b** **c**

Figure 2.14. (a) When evaluated by the IoU between the detection bounding-box and the ground-truth bounding-box, SCRL achieves great overlap on all kind of VBs. (b) When evaluated by Loc-Err(namely, the centroid distance between the detection and the ground-truth), SCRL accurately detects all kind of VBs. (c) SCRL achieves a great VB identification result. SCRL achieves 96% IDR when the identification threshold is set to 3 mm.

and regress the bounding-box accurately. For the middle VBs in the lumbar spine (i.e., L5-L1), the average IoU for these VBs is higher than the sacrum by 2%. In Fig. 2.14(b), the average Loc-Err for these VBs is lower than T10 by 1 pixel. Such a better result is mainly because these VBs are spatially correlated to previous adjacent VBs. Thus, AMRL models this anatomical correlation to focus the attention-region more accurately for these VBs. With more accurate attention-region providing optimal analysis view, FC-ResNet is able to achieve a higher detection accuracy [23]. For the sacrum, SCRL achieves 90% average IoU and 2.7-pixel mean Loc-Err. This excellent result indicates that although the shape and data-size of the sacrum are quite different and smaller compared with the other lumbar VBs, SCRL is robust to manage these challenges.

Based on the IDR result under various identification threshold(Fig. 2.14(c)), SCRL totally satisfies the condition of accurate VB detection. SCRL achieves 96% IDR when the identification threshold is set to 3 mm. The above excellent performance fully demonstrates that SCRL achieves a great VB identification result and has the ability to obtain the detailed representation of VBs and achieves accurate VB detection result to assist clinicians for spine analysis.

Comparison with existing methods. As shown in Table. 2.2 and Fig. 2.15, the performance of SCRL is overall much better than these existing methods. SCRL significantly outperforms TDCN [13] by 10.50% in mIoU, 4.152 pixels in Loc-Err, and 11.9% in IDR when the identification threshold is 3 mm. TDCN gets 81.8% in IoU, 7.794 pixels in Loc-Err, and 84.4% in IDR. Also, SCRL outperforms Hi-scene [9] by 0.4% in mIoU, 2.057 pixels in Loc-Err. Hi-scene gets 91.9% in mIoU, 5.199 pixels in Loc-Err and 96.3% in IDR. The above data proves SCRL is more capable of excluding the influence (such as the fat tissue having a similar appearance to VBs) and modeling the various representation of different VBs (such as scale variations and pathological variations) to detect the accurate location of VBs. A visualized example is the detection result of above methods for the patient 1 in Fig. 2.15, where the L5 VB is deformable and SCRL's detection is the best compared with other methods' detection.

Effectiveness of FC-ResNet for detection. Benefiting from the attention-region proposal by the AMRL, FC-ResNet further improves the detection accuracy effectively and achieves great detection performance. According to Table. 2.3, the attention-region by AMRL achieves 83.6% in IoU, 4.735 pixels in Loc-Err, and 91.7% in IDR. Based on the above coarse detection, FC-ResNet further increases IoU by 8.7%, decreases Loc-Err by 1.593 pixels, and increases IDR by 4.6%. The above increased detection accuracy demonstrates the effectiveness of FC-ResNet for VB detection. AMRL proposes coarse location and size of the desirable VB by modeling spine anatomy, FC-ResNet predicts more accurate VB location and size on the basis AMRL,

where the detection improvement between AMRL and FC-ResNet is displayed in Fig. 2.15.

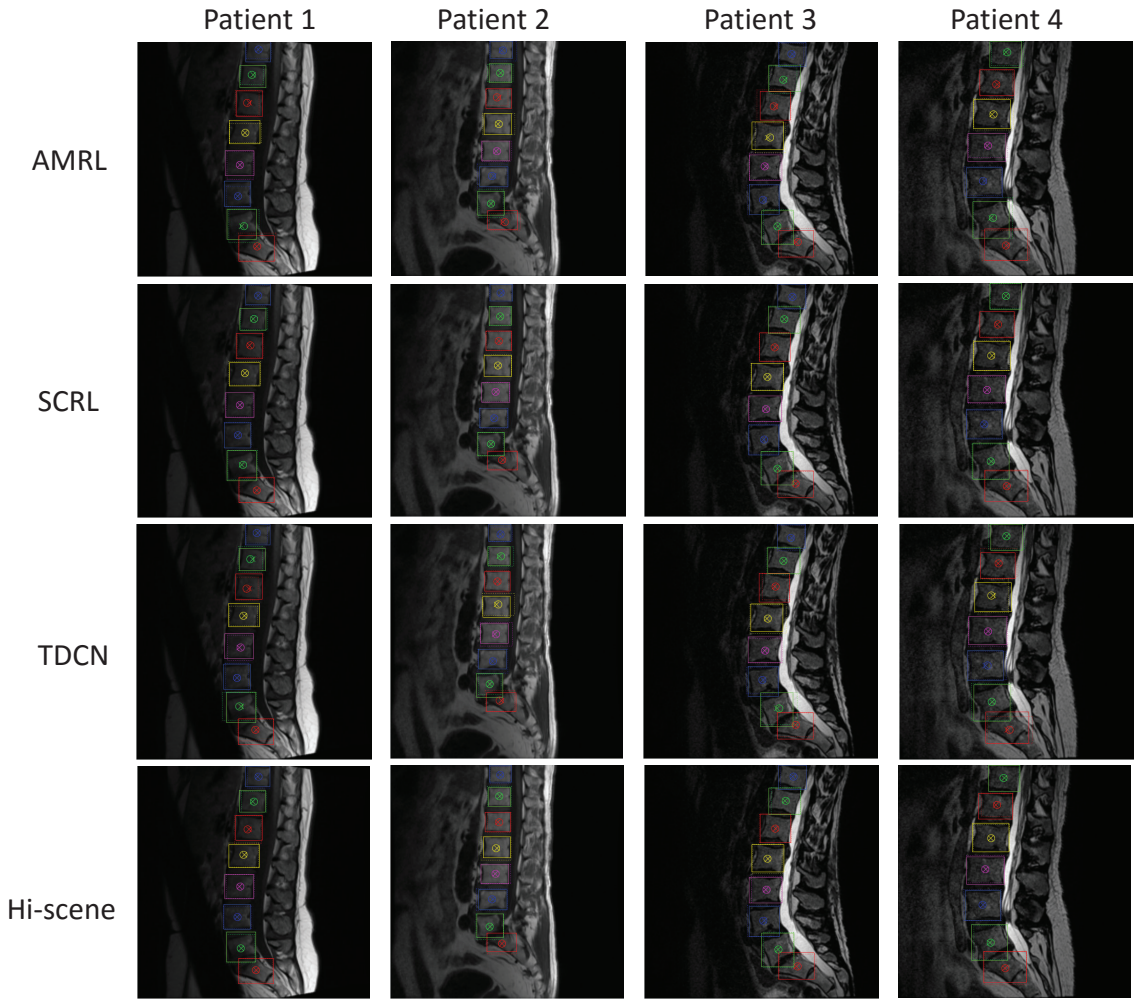


Figure 2.15. Some visualization examples of VB detection, where the solid box and circle indicate the bounding-box and centroid of every VB from the ground-truth, the dotted box and ”x” indicate the bounding-box and centroid of every VB from the prediction of every methods. Based on the overlap of predicted bounding-box to the ground-truth bounding-box and the distance of the predicted centroid to the ground-truth centroid, the performance of SCRL is best, especially when the VB is deformed. In addition, the high-overlapping between the attention-region of AMRL and the ground-truth indicates the effectiveness of AMRL to model the spine anatomy and transform the attention-region from the previous VB to the next VB.

Advantages of FC-ResNet compared to the segmentation-based two-step method in VB detection. FC-ResNet is more advantageous to predict the bounding-box compared with the segmentation-based two-step method (i.e., obtaining the surrounding from the segmentation mask by Y-Net). According to Table. 2.3, benefiting from the AMRL, both the bounding-

Table 2.3. The detection performance among the attention-region by AMRL, the surrounding of segmentation mask by Y-Net, and the bounding-box by FC-ResNet.

Methods	IoU	Loc-Err	IDR ($\tau = 3mm$)
AMRL	0.836	4.735	0.917
AMRL+Y-Net (surrounding)	0.914	3.268	0.951
AMRL+FC-ResNet (bounding-box)	0.923	3.142	0.963

box by FC-ResNet and the surrounding by Y-Net achieve great detection accuracy. However, because of the cumulative error, the accuracy of the surrounding is a little lower than the accuracy of the directly-predicted bounding-box. Considering more accurate VB detection provides a wider range of clinical assistance, such as measuring the narrowing level of the spinal canal when diagnosing spinal stenosis[41], the SCRL reserves the FC-ResNet to predict the bounding-box instead of taking the surrounding of the segmentation mask as the detection result.

Segmentation result

Qualitative evaluation. SCRL completes an excellent segmentation despite the complex appearance and various spatial offsets of VBs. In Fig. 2.16, the segmentation visualization results look reasonable and the boundary of each VB region was extracted nicely compared with the ground-truth. Even there are some serious appearance changes caused by vertebrae diseases, such as spondylolisthesis and vertebral collapse, SCRL effectively segments these spine images accurately. For instance, in Fig. 2.16(c) the shape of L5 has changed because of spondylolisthesis, SCRL is robust to this geometry variation and segments these VBs accurately. L5 in Fig. 2.16(b)(d) are deformed because of the vertebral fracture, SCRL captures this small difference and presents these deformations clearly on the segmentation mask.

Quantitative analysis. Based on the first row of Table. 2.2, SCRL achieves significant segmentation accuracy. The first row of Table. 2.2 quantitatively demonstrates the effectiveness of the segmentation result of SCRL. Overall SCRL achieves 94.9% in PA and 92.6% in Dice, which means SCRL has a strong ability to classify each VB pixel correctly from the background. SCRL achieves 1.919mm in HD, which means SCRL is capable of finding the true VB boundaries despite the complex VB shape and geometry.

Based on the segmentation result of each kind of VB(Fig. 2.17(a)(b)(c)), SCRL has the ability to cope with all kinds of VBs and segment them accurately. For the first VBs (such as T10-T12), the average PA, Dice, and HD of SCRL segmentation in these VBs are over 95%, 93%,

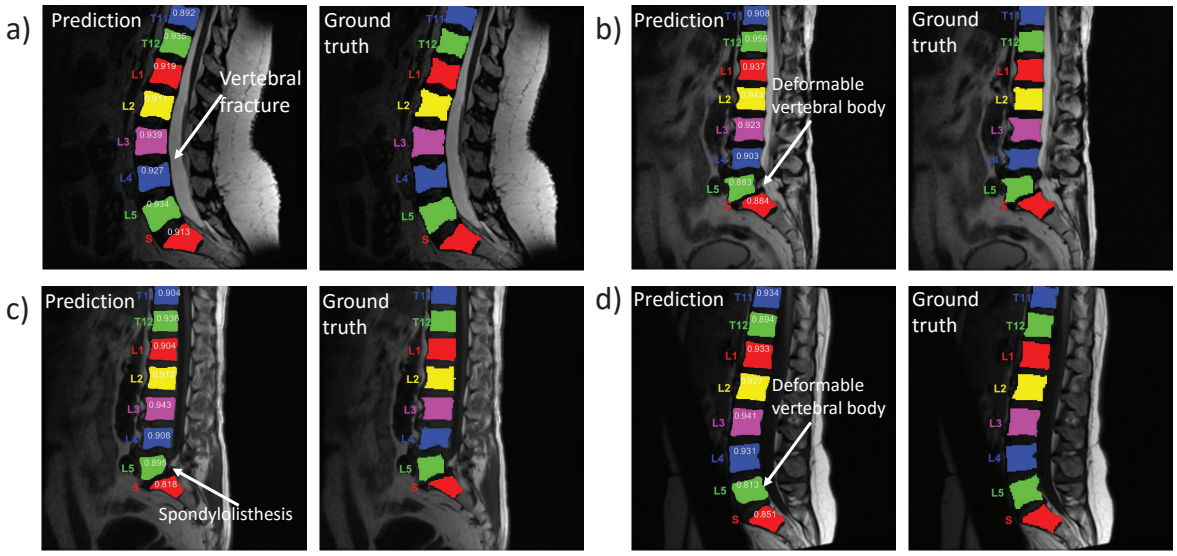
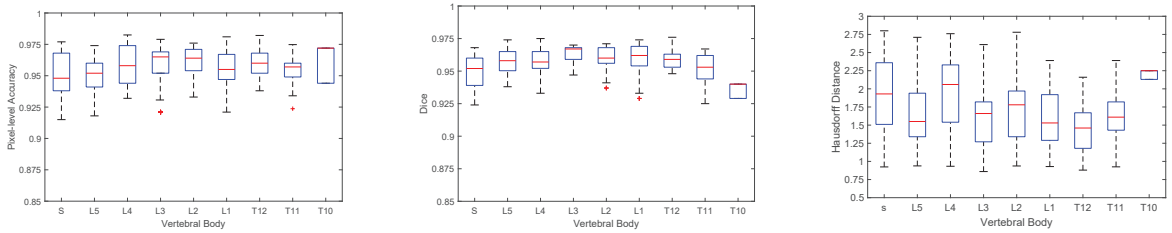


Figure 2.16. Some visualizations of VB segmentation results by SCRL. Although there are some appearance differences caused by vertebrae diseases, such as spondylolisthesis, vertebrae collapse and vertebral fracture, SCRL is robust to these geometry variations and presents these deformations clearly on the segmentation mask.

and below 2.5mm respectively. Such great performance indicates AMRL successfully searches and focuses the attention-region on them and then Y-Net segments them accurately. For these middle lumbar VBs, such as L5-L1, SCRL achieves better segmentation results. For instance, in Fig. 2.17(a)(b), the average PA and Dice for these VBs are higher than the sacrum by 1.5%. In Fig. 2.17(c), the average HD for these VBs is lower than T10 by 0.5 mm. Such great improvement is because there are adjacent VBs for AMRL to model the spatial correlation to focus the attention-region accurately. Thus, better attention-region provides Y-Net with a better analysis view, which leads to more precise segmentation. For the sacrum, SCRL successfully



a **b** **c**

Figure 2.17. (a) When evaluated by the PA of correctly classified VB pixel, SCRL achieves high true-positive classification for each kind of VBs. (b) When evaluated by Dice, SCRL obtains great overlap between the segmentation mask and ground-truth. (c) When evaluated by HD, SCRL gets significant matching between the segmentation contour and the ground-truth contour.

Table 2.4. Segmentation performance comparison among AMRL+U-Net, AMRL+Y-Net, and a separate U-Net

Methods	PA	Dice	HD
AMRL+U-Net	0.951	0.925	1.731
AMRL+Y-Net	0.949	0.926	1.943
U-Net	0.930	0.914	2.363

copes with the different appearance and segments it accurately. According to Fig. 2.17(a)(b)(c), the segmentation result gets average PA, Dice, and HD of SCRL segmentation in the sacrum over 93%, 95%, and 2.0mm respectively. Thus, although the appearance of sacrum is totally different from other lumbar VBs, SCRL effectively learns the texture feature from multi-scales and outputs an accurate segmentation.

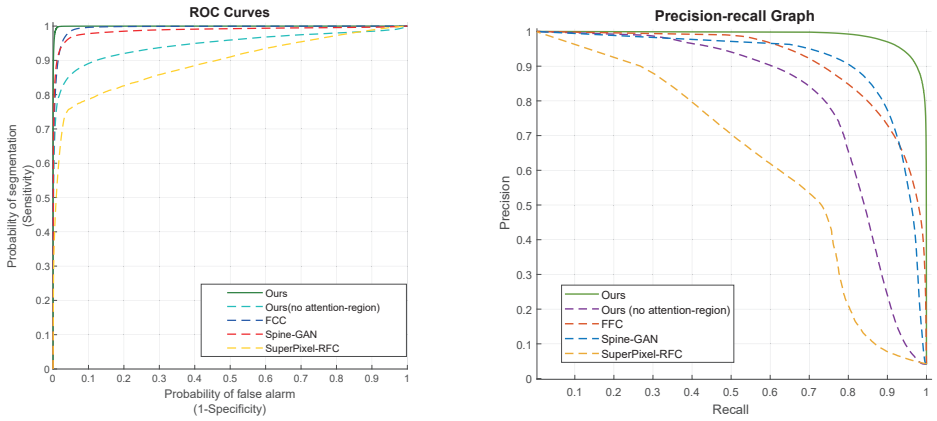


Figure 2.18. ROC Curve and Precision-recall Graph for SCRL and other methods. The area under the curve of SCRL is much bigger than the area under the other method curves, which indicates SCRL has a stronger ability in classifying the positive pixel as VBs rather than classifying the negative pixel as VBs.

The influence of classification branch to segmentation. To obtain the influence of classification branch to segmentation, the experiment deploys a U-Net parallel to the Y-Net (without classification branch) after AMRL. According to Table. 2.4, AMRL + U-Net achieves 95.1% PA, 92.5% in Dice, and 1.731 mm in HD. Benefit from the attention-region by AMRL, the segmentation performance of AMRL+Y-Net and AMRL+U-Net is overall comparable. Particularly, the classification branch reduces the perception of Y-Net to the VB edge, which causes the HD increasing and mPA increasing. The classification branch also makes Y-Net conservative, which causes the Dice increasing. Compared with the AMRL+U-Net, the segmentation performance of Y-Net does not deteriorate but improves in certain aspects by combining the segmentation and classification. In addition, considering the U-Net takes more GPU memory,

Y-Net combining segmentation and classification into a single network is a better choice for SCRL.

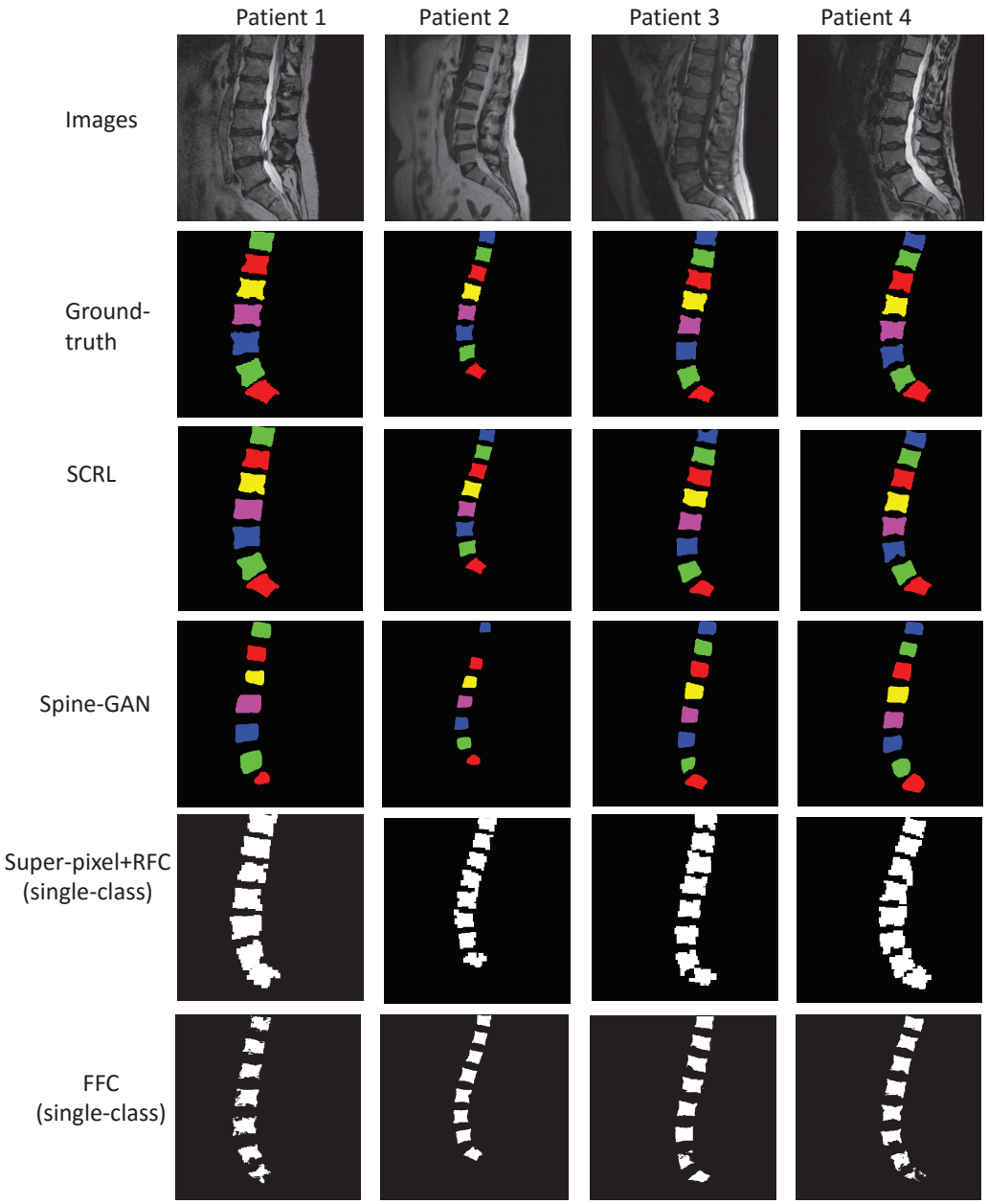


Figure 2.19. Some visualization examples of VB segmentation. Note that, the segmentation of FFC and Super-pixel+RFC is binary. After comparing with the results of other methods, the result of SCRL is the best. In detail, the FFC cannot handle the size-variation of VBs, which can be observed from the segmentation of patient 1 and patient 4. The Spine-GAN achieves great instance segmentation result, but it misses the T12 VB in patient 2. The SuperPixel-RFC overall performs better than FFC, but it leverages super-pixels, which causes it cannot handle the VB boundary area effectively.

The effectiveness of attention-regions to segmentation. To obtain the effectiveness of attention-region to segmentation, the experiment also deploys a separate U-Net (independent from SCRL). According to Table. 2.4, The separate U-Net achieves 93.0% PA, 91.4% in Dice, and 2.363 mm in HD. According to this result, we find that without the attention-region as a priori, the segmentation performance decreases in every aspect. In detail, compared with the result of Y-Net, the mPA decreases by 1.9%, the Dice decreases by 1.2%, and the HD increases by 0.632 mm. Thus, the attention-region is important for an accurate VB segmentation. In addition, the segmentation result of the separate is binary segmentation instead of instance segmentation, which also reduces its comparability to our Y-Net.

Comparison with existing methods. Based on the Table. 2.2, SCRL achieves more accurate segmentation results compared with the other three existing methods. SCRL significantly outperforms FFC by 4.4% in PA, 3.6% in Dice, and 0.619 mm in HD. FFC gets 90.5% in PA, 89.0% in Dice, and 2.362 mm in HD. Also, SCRL outperforms Spine-GAN by 2.7% in PA, 0.9% in Dice, and 0.663 mm in HD. Spine-GAN gets 92.2% in PA, 91.7% in Dice, and 2.406 mm in HD. SCRL outperforms SuperPixel-RFC by 7.0% in PA, 8.5% in Dice, and 2.517 mm in HD. SuperPixel-RFC gets 87.9% in PA, 84.1% in Dice, and 4.260 mm in HD. Hence, SCRL has a stronger perception of the appearance of the VBs and models the VB surroundings more accurately to distinguish the true VB from surrounding tissues.

In addition, SCRL is stronger than other VB segmentation methods for pixel classification. The AUC (Area Under the Curve) of SCRL is much bigger than AUCs of other methods in the ROC Curves and Precision-recall Graph (Fig. 2.18). This means SCRL has a robust ability that classifies the positive pixel (the pixel belongs to VBs) as VBs rather classifying the negative pixel (the pixel belongs to the background) as VBs.

Classification result

Qualitive analysis. Table. 2.2 SCRL achieves excellent classification performance, where SCRL gets 96.4% in mean accuracy and 95.8% in image accuracy. The former indicates that 96.4% VBs are classified correctly, and the latter indicates that all VBs are classified correctly in 95.8% images. By checking the result, we find only one image is mis-labeled in the testing dataset. In this mis-labeled image, the classification branch mis-classifies a sacrum as a lumbar VB, which further causes all the VB in this spine image are mis-labeled. We analyze this is because the imbalance between the number of the lumbar VB and the number of the sacrum, which leads to the classification branch over-fitting.

Comparison with existing methods. As shown in Table. 2.2, the classification performance of SCRL is overall better than these existing methods. SCRL outperforms TDCN 12.5% in image accuracy and 4.8% in mean accuracy, where TDCN only achieves 83.3% in image accuracy. Compared with another state-of-art method Hi-scene, the image accuracy of SCRL is higher than it by 2.5%. Hi-scene employs the mask-rcnn architecture, thus its classification ability is strong enough as a baseline. The out-performance of SCRL to Hi-scene indicates SCRL is quite competitive in classification. Compared with Spine-GAN, SCRL outperforms it 4.1% in image accuracy, which illustrates the spine-anatomy-modeling of SCRL facilitates SCRL has a stronger classification ability than Spine-GAN that employs a encode-decode architecture to model spine spatial features.

Effectiveness of AMRL for attention-focusing

Action-related hyperparameters tuning. Based on the hyperparameter tuning experimental result in Tab. 2.5, we find when the number of actions is set as 20 and step-sizes are set as 6, 4, the attention-region accuracy was the highest, which is 83.6%. Considering the action prediction in Soft Actor-Critic (SAC) is symmetric about zero, we further set the step-sizes of displacement and size-adjustment as [-6,6] and [-4,4] respectively. This setting is a balance of other two settings. In detail, for the setting that number of actions is 15, the attention-focusing accuracy is only 76.7%, where the small number of actions restricts the agent-exploration to the image, and the large step-sizes also increases the action-prediction difficulty. For the setting that number of actions is 15, the attention-focusing accuracy is 81.4%, where the large number of actions causes too many explorations and thus may cause over-fitting of the decision-making.

Table 2.5. The attention-focusing results under different action-related hyperparameter-settings

Action-related hyperparameters		Attention-region accuracy
Number of actions	step-size of action $\{displacement,\ size\ adjustment\}$	
15	$\{8, 6\}$	0.767
20	$\{\mathbf{6}, \mathbf{4}\}$	**0.836**
25	$\{5, 3\}$	0.814

* Number of actions × Step-size of action ≥ Max task demands(Max displacement demand = 106, Max size-adjustment demand = 67). For instance, 15×8=120>106, 15*6=80>67.

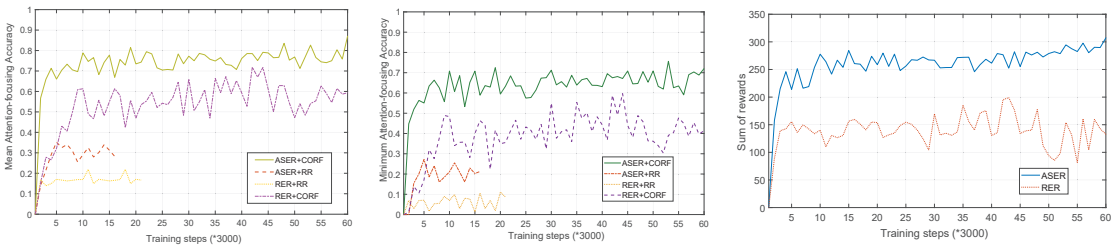
Effectiveness of ASER and CORF for attention-focusing. According to the experiment results in Fig. 2.20, the ASER and CORF have respective contributions in attention-focusing accuracy and training stability. In detail:

- AMRL has the ability to focus attention-regions on each VB. In Fig. 2.20(a)(b), AMRL achieves 84.00% mean attention-focusing accuracy and 75.00% minimum attention-focusing accuracy in the validation dataset during training. This excellent attention-focusing result demonstrates AMRL's ability to approach each desirable VB despite the critical scale variations and pathological variations of VBs, and effectively propose accurate attention-regions.

- CORF successfully stimulates AMRL to achieve higher attention-focusing accuracy. Compared with the regular reward function(RR), as shown in Fig. 2.20(a)(b), CORF improves the mean and minimum attention-focusing accuracy by 30% respectively. Such significant accuracy improvement indicates that the exponential weight of CORF is really beneficial in encouraging the agent network to select positive actions so as to increase the attention-focusing accuracy.

- ASER effectively promotes attention-focusing accuracy. Compared with the regular experience-replay (RER), ASER improves the mean and minimum attention-focusing accuracy by 10% and 15% respectively(Fig. 2.20(a)(b)). In addition, ASER obtains almost twice the sum of rewards than RER(Fig. 2.20(c)) when they are under the same reward function. This great improvement demonstrates the unique experience-sampling method of ASER (namely, ASER increases the training-data difficulty as the competency of the agent network grows) increases the utilization efficiency of the history experience and the learning efficiency of the agent network. As a result, ASER increases the decision-making ability of the agent network, thereby enabling the agent network to select correct actions to increase the attention-focusing accuracy.

- ASER enhances the stability of AMRL compared with the regular experience-replay(RER).



(a) **(b)** **(c)**

Figure 2.20. (a)(b) shows the attention-focusing accuracy of cross-combination experiments (namely, i. ASER+CORF; ii. ASER+regular reward function (RR); iii. regular experience-replay(RER)+CORF; iv. RR+RER) in validation dataset during training period; (c) illustrates the reward obtained by ASER and regular experience-replay (RR) under CORF in validation dataset during training period.

In Fig. 2.20(a)(b)(c), the fluctuation frequency and amplitude of the curve with ASER is much less than the fluctuation of the curve with RER. For instance, in Fig. 2.20(c), the fluctuation amplitude of the ASER curve is less than 50, while the fluctuation amplitude of the RER curve is more than 75. This stability enhancement indicates the experience-sampling method is meaningful for gradually and stably improving the decision-making ability of the agent network.

References

[1] S. V. Kushchayev, T. Glushko, M. Jarraya, K. H. Schuleri, M. C. Preul, M. L. Brooks, and O. M. Teytelboym, "Abcs of the degenerative spine," *Insights into imaging*, vol. 9, no. 2, pp. 253–274, 2018.

[2] D. Štern, B. Likar, F. Pernuš, and T. Vrtovec, "Parametric modelling and segmentation of vertebral bodies in 3d ct and mr spine images," *Physics in Medicine & Biology*, vol. 56, no. 23, p. 7505, 2011.

[3] E. McCloskey, H. Johansson, A. Oden, and J. A. Kanis, "Fracture risk assessment," *Clinical biochemistry*, vol. 45, no. 12, pp. 887–893, 2012.

[4] G. Tatoń, E. Rokita, M. Korkosz, and A. Wróbel, "The ratio of anterior and posterior vertebral heights reinforces the utility of dxa in assessment of vertebrae strength," *Calcified tissue international*, vol. 95, no. 2, pp. 112–121, 2014.

[5] M. Rak, J. Steffen, A. Meyer, C. Hansen, and K.-D. Tönnies, "Combining convolutional neural networks and star convex cuts for fast whole spine vertebra segmentation in mri," *Computer Methods and Programs in Biomedicine*, 2019.

[6] J. Yao, J. E. Burns, H. Munoz, and R. M. Summers, "Detection of vertebral body fractures based on cortical shell unwrapping," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 509–516, Springer, 2012.

[7] B. CARLSON, "Scoliscore ais prognostic test personalizes treatment for children with spinal curve," *Biotechnology healthcare*, vol. 8, no. 2, p. 30, 2011.

[8] M. W. Smith, J. Reed, R. Facco, T. Hlaing, A. McGee, B. M. Hicks, and M. Aaland, "The reliability of nonreconstructed computerized tomographic scans of the abdomen and pelvis in detecting thoracolumbar spine injuries in blunt trauma patients with altered mental status," *JBJS*, vol. 91, no. 10, pp. 2342–2349, 2009.

[9] S. Zhao, X. Wu, B. Chen, and S. Li, "Automatic vertebrae recognition from arbitrary spine mri images by a hierarchical self-calibration detection framework," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 316–325, Springer, 2019.

[10] Y. Cai, M. Landis, D. T. Laidley, A. Kornecki, A. Lum, and S. Li, "Multi-modal vertebrae recognition using transformed deep convolution network," *Computerized medical imaging and graphics*, vol. 51, pp. 11–19, 2016.

[11] B. Gaonkar, Y. Xia, D. S. Villaroman, A. Ko, M. Attiah, J. S. Beckett, and L. Macyszyn, "Multi-parameter ensemble learning for automated vertebral body segmentation in heterogeneously acquired clinical mr images," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–12, 2017.

[12] M. Rak and K.-D. Tonnies, "A learning-free approach to whole spine vertebra localization in mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 283–290, Springer, 2016.

[13] Y. Zhou, Y. Liu, Q. Chen, G. Gu, and X. Sui, "Automatic lumbar mri detection and identification based on deep learning," *Journal of digital imaging*, vol. 32, no. 3, pp. 513–520, 2019.

[14] J.-T. Lu, S. Pedemonte, B. Bizzo, S. Doyle, K. P. Andriole, M. H. Michalski, R. G. Gonzalez, and S. R. Pomerantz, "Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning," *arXiv preprint arXiv:1807.10215*, 2018.

[15] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, "Spine-gan: Semantic segmentation of multiple spinal structures," *Medical image analysis*, vol. 50, pp. 23–35, 2018.

[16] A.-A.-Z. Imran, C. Huang, H. Tang, W. Fan, K. Cheung, M. To, Z. Qian, and D. Terzopoulos, "Analysis of scoliosis from spinal x-ray images," *arXiv preprint arXiv:2004.06887*, 2020.

[17] J. Yi, P. Wu, Q. Huang, H. Qu, and D. N. Metaxas, "Vertebra-focused landmark detection for scoliosis assessment," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 736–740, IEEE, 2020.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[19] F.-C. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu, "Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 176–189, 2017.

[20] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Young Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2711–2720, 2017.

[21] C. Huang, S. Lucey, and D. Ramanan, "Learning policies for adaptive tracking with deep feature cascades," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 105–114, 2017.

[22] S. Mathe, A. Pirinen, and C. Sminchisescu, "Reinforcement learning for visual object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2894–2902, 2016.

[23] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9080–9089, 2018.

[24] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[25] C. Xu, L. Xu, Z. Gao, S. Zhao, H. Zhang, Y. Zhang, X. Du, S. Zhao, D. Ghista, H. Liu, *et al.*, "Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture," *Medical image analysis*, vol. 50, pp. 82–94, 2018.

[26] R. Zhao, W. Liao, B. Zou, Z. Chen, and S. Li, "Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis," 2019.

[27] C. Xu, J. Howey, P. Ohorodnyk, M. Roth, H. Zhang, and S. Li, "Segmentation and quantification of infarction without contrast agents via spatiotemporal generative adversarial learning," *Medical Image Analysis*, p. 101568, 2019.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[30] D. H. Hubel and T. Wiesel, "Shape and arrangement of columns in cat's striate cortex," *The Journal of physiology*, vol. 165, no. 3, pp. 559–568, 1963.

[31] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[32] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu, "An artificial agent for robust image registration," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[33] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao, and A. Kamen, "Robust non-rigid registration through agent-based action learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 344–352, Springer, 2017.

[34] A. Alansary, L. Le Folgoc, G. Vaillant, O. Oktay, Y. Li, W. Bai, J. Passerat-Palmbach, R. Guerrero, K. Kamnitsas, B. Hou, *et al.*, "Automatic view planning with multi-scale deep reinforcement learning agents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 277–285, Springer, 2018.

[35] 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.

[36] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Išgum, "Iterative fully convolutional neural networks for automatic vertebra segmentation and identification," *Medical image analysis*, vol. 53, pp. 142–155, 2019.

[37] L. Lu, Y. Zheng, G. Carneiro, and L. Yang, "Deep learning and convolutional neural networks for medical image computing," *Advances in Computer Vision and Pattern Recognition*, 2017.

[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[40] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 590–598, Springer, 2012.

[41] J. Steurer, S. Roner, R. Gnannt, and J. Hodler, “Quantitative radiologic criteria for the diagnosis of lumbar spinal stenosis: a systematic literature review,” *BMC musculoskeletal disorders*, vol. 12, no. 1, p. 175, 2011.

CHAPTER 3

The contents of this chapter was previously submitted in the Medical Imaging Analysis.

3 DRL-BASED WEAKLY-SUPERVISED TEACHER-STUDENT NETWORK FOR LIVER TUMOR SEGMENTATION WITHOUT CONTRAST AGENT

3.1 Introduction

Accurate liver tumor segmentation is indispensable for clinicians to diagnose and treat the liver tumor and improve the survival rate of patients [1]. 1) The liver tumor segmentation assists clinicians to decide the stage of the tumor and make the liver surgical planning (e.g., oncologic resections and liver transplantation). For instance, current all commonly used staging systems, such as the Barcelona Clinic Liver Cancer staging system, take into account tumor size as well as lesion multiplicity to select suitable candidates for surgical treatment or local-regional therapies [2]. 2) The liver tumor segmentation assists clinicians to simplify the planning for surgical resection. One of the main constraints for surgical resection planning is the lesion/liver ratio after surgical resection[3]. The identification of the regions to be removed becomes easier as tumors are well defined, the segmentation also provides a precise location of the tumors inside the anatomical segments of the liver simplifying the preoperational planning[4]. 3) The liver tumor segmentation assists clinicians to track the treatment response and evaluate the therapy effect on tumors [5]. The quantitative assessment of tumor segmentation (e.g., volume and diameter) helps physicians to determine whether the therapy is effective, which improves the rehabilitation rate of patients and reduces the medication costs.

Almost all existing tumor segmentation methods are only suitable for enhanced images where a contrast agent (CA) is used to improve tumor visibility (Fig. 3.1) and thus raises several issues. 1) The high risk caused by the potential toxicity of CAs [6, 7]. For instance, about 1% gadolinium-based CA is retained in the tissues after injection, which may cause 10%-15% of the incidence of Contrast-Induced Nephropathy (CIN) [8]. 2) The time-consuming caused by the CA injection and the imaging process after the injection. For instance, gadolinium dimeglumine (Gd-BOPTA, MultiHance) requires 40-120 mins to point focal hepatic lesions out as dark lesions in contrast to the enhancing normal liver [9]. 3) The expensive cost caused by CA materials and image scanning. CA materials will be wasted if errors occur during the injection. In addition, to obtain the enhanced image, it also needs a second scanning (namely Dynamic Contrast-Enhanced Magnetic Resonance Imaging) and requires extra cost.

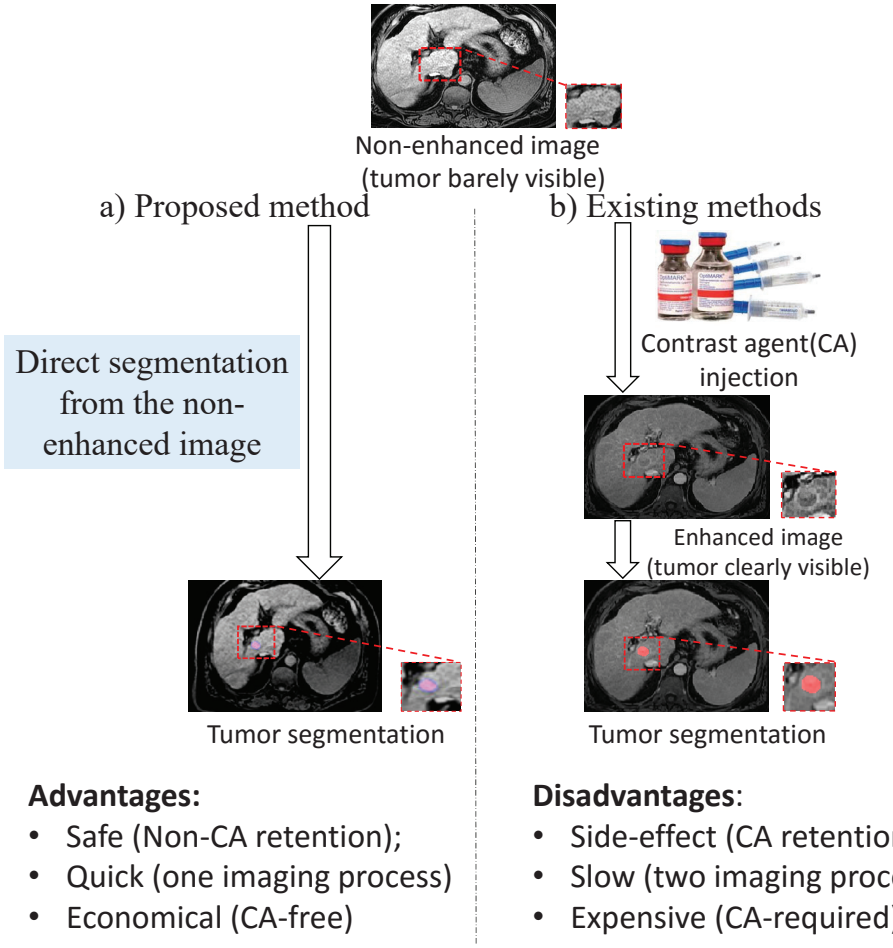


Figure 3.1. Almost all existing tumor segmentation methods are only suitable for enhanced images where a contrast agent (CA) is used to improve tumor visibility and thus raises several issues (side-effect, time-consuming, and high cost). We propose a method to segment the liver tumor from non-enhanced images, which is safe, quick, and economical because it avoids the use of CAs.

Liver tumor segmentation from non-enhanced images can avoid the above disadvantages caused by the CA, but it is challenging (Fig. 3.2): 1) Some tumors are barely visible to naked eyes in the non-enhanced image, which usually causes that these tumors cannot be identified accurately and segmented from non-enhanced images. Without contrast agents enhancing, the contrast between the tumor and normal tissues is really low. Apparent tumor features thus are difficult to be extracted from non-enhanced images to segment the liver tumor. 2) Normal tissues in the liver area present a complex and tumor-like appearance, which often causes traditional Convolutional Neural Network (CNN) based methods to get false-positive segmentation in the normal tissues. Because tumor features produced under the inherent down-sampling mechanism are of low resolutions commonly, the strong feature of normal tissues is reserved

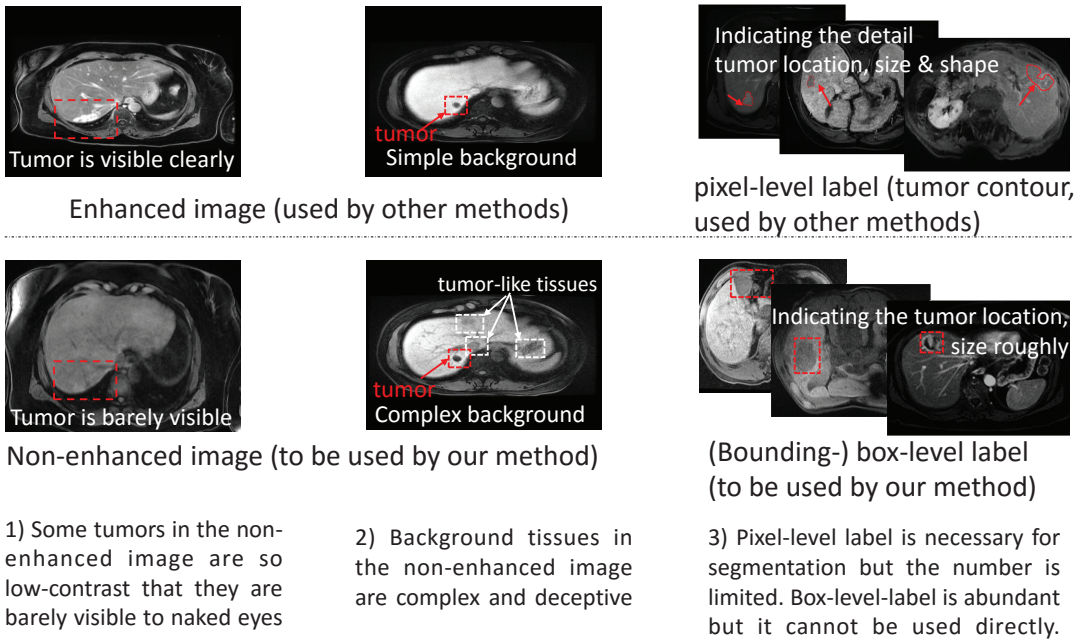


Figure 3.2. Tumor segmentation from non-enhanced images facing the following challenges: 1) Some tumors are barely visible to naked eyes. 2) Normal tissues in the background present complex and tumor-like appearance and thus may cause interference to the tumor segmentation. 3) The pixel-level-labeled data is limited and the box-level-labeled data is relatively abundant. It is desirable but challenging to deploy the box-level-labeled data for tumor segmentation training.

and the weak feature of tumors is lost. Therefore, in traditional CNN-based methods, the normal tissue is possibly classified as the tumor incorrectly. 3) Accurate pixel-level-labeled data is limited, it is highly-desirable to leverage (bounding-)box-level-labeled data to improve segmentation performance. The tumor is barely visible in the non-enhanced image and requires deep networks to extract valuable features. Therefore, the deep network involves a huge number of parameters and requires a decent amount of training data [10]. However, it is expensive and tedious to delineate reliable tumor contours by experienced experts, especially for those tumors with small size and complex shapes. While the (bounding-)box-level-labeled data is relatively easy to obtain because the label only indicates the tumor location and size. It is helpful but challenging to exploit box-level-labeled data for raising tumor segmentation performance.

Although RgGAN [11] has been proposed to segment the tumor from the non-enhanced image, it cannot address the above challenges. RgGAN inputs the radiomics-features in contrast images as additional knowledge into the discriminator to regulate the feature extraction of the non-contrast image in the generator (segmenter). The radiomics-feature promotes the perfor-

mance of discriminator, but the regulation from discriminator is too limited (a binary value as feedback). Thus the radiomics-features have little effect on the generator (segmenter) for pixel-level segmentation. In addition, RgGAN is a fully-supervised method and requires a reasonable number of pixel-level-label data to support the training. It cannot make use of the box-level-labeled data.

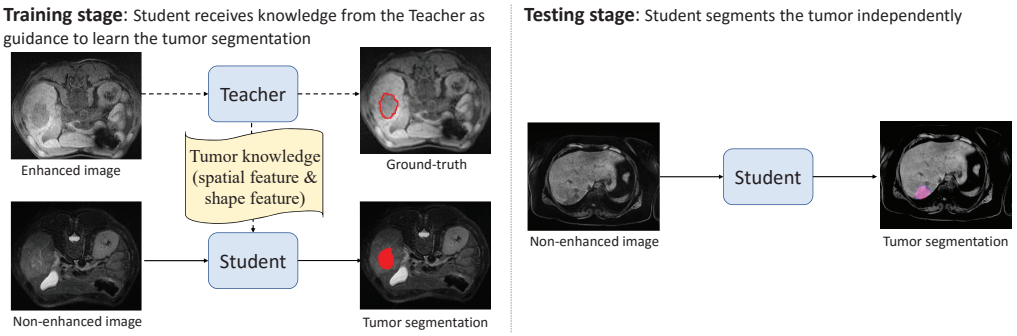


Figure 3.3. The teacher-student framework draws support from the enhanced image where the tumor is high-contrast and clearly visible to enable the tumor segmentation in the non-enhanced image. By deploying a teacher to impart the tumor knowledge (including tumor spatial and shape features) from the enhanced image to a student in the non-enhanced image, the teacher-student framework exploits the enhanced image to guide the tumor segmentation learning in the non-enhanced image during the training stage. Thus during the testing stage, the student is able to segment the tumor independently from the non-enhanced image.

Transferring tumor knowledge from the enhanced image as guidance has great potential to address the tumor segmentation in the non-enhanced image. Namely, employing a teacher-student framework to facilitate the tumor segmentation in the non-enhanced image. As shown in Fig. 3.3, in the teacher-student framework, a teacher in the enhanced image learns and imparts the tumor knowledge (tumor spatial and shape feature) to a student in the non-enhanced image. The tumor knowledge facilitates the student to deal with the low-contrast tumor and complex backgrounds, thus the student is able to segment the tumor independently from the non-enhanced image during the testing stage. However, there are **two issues** in such teacher-student framework: 1) The data distribution differs between the enhanced and non-enhanced image, i.e., these two kinds of images belong to each different domains. This distribution difference causes it is difficult for the teacher to select suitable tumor spatial feature in the enhanced image to assist the student to identify and detect the tumor in the non-enhanced image. 2) The box-level-labeled data lacks detailed tumor shape information because the label only indicates the tumor location and size. Therefore, for the box-level-labeled data, it is hard to extract accurate tumor shape features from the enhanced image to improve (instead of misleading) the tumor segmentation effectively in the non-enhanced image.

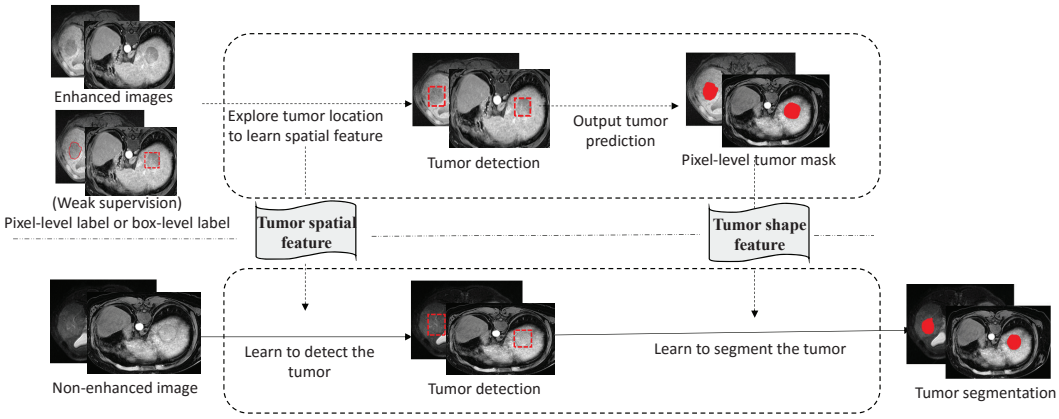


Figure 3.4. WSTS addresses the weakly-supervised tumor segmentation from the non-enhanced image by taking advantage from the enhanced image during the training stage. In detail, WSTS explores the tumor location to learn tumor spatial feature and predicts an accurate pixel-level tumor mask for the box-level-labeled data as tumor shape feature, thus taking the tumor spatial and shape features as guidance, WSTS learns to detect and segment the tumor in the non-enhanced image.

In this paper, we propose a Weakly-Supervised Teacher-Student network (WSTS, Fig. 3.4) to address the liver tumor segmentation in the non-enhanced image by additionally leveraging the box-level-labeled data. To this purpose, WSTS employs a weakly-supervised teacher-student framework (TCH-ST). In detail, in the training stage, WSTS explores the tumor location to learn tumor spatial feature in the enhanced image and predicts an accurate pixel-level tumor mask for the box-level-labeled data as tumor shape feature. With the tumor spatial and shape features as guidance, WSTS learns to detect and segment the tumor in the non-enhanced image. Thus in the testing stage, WSTS is able to detect and segment the tumor from the non-enhanced image without the assistance of the enhanced image. To determine the tumor location and size correctly (the detail motivation see Sec. 3.2.1), WSTS proposes *Dual-strategy DRL (DDRL)*. The DDRL develops two tumor detection strategies to jointly determine the tumor location in the enhanced image by creatively introducing a relative-entropy bias in the DRL. By following the detection strategies, WSTS is able to determine the tumor location in the non-enhanced image. To predict the tumor mask precisely for the box-level-labeled data (the detail motivation see Sec. 3.2.2), WSTS proposes an *Uncertainty-Sifting Self-Ensembling (USSE)*. The USSE utilizes the limited pixel-level-labeled data and additional box-level-labeled data to predict the tumor accurately by evaluating the prediction reliability with a newly-designed Multi-scale Uncertainty-estimation. By taking the tumor prediction as a pseudo label (additional to the manual pixel-level-label), the tumor segmentation in the non-enhanced image is thus improved.

Our main contributions can be summarized as follows:

- The new weakly-supervised teacher-student framework (TCH-ST) we proposed integrates Deep Reinforcement Learning and Self-Ensembling, which for the first time achieves liver tumor segmentation from the non-enhanced image by exploiting the additional (bounding-)box-level-labeled data. This provides a new solution for exploiting different imaging modalities in a weakly-supervised manner, and using machine learning to retain the same level of information from the less risky one.

- For the first time, the Dual-strategy DRL (DDRL) develops two detection strategies to locate the tumor jointly, which increases the DRL exploration range in the image and avoids the situation that traditional DRL (single strategy) sticks into sub-optimal and causes inaccurate tumor detection. To this purpose, DDRL adds a relative-entropy bias in the DRL learning objective, which provides a novel solution for applying multiple-strategies in object detection.

- The Uncertainty-Sifting Self-Ensembling (USSE) improves the tumor prediction reliability in the enhanced image by integrating uncertainty-estimation with Self-Ensembling, which prevents the error magnifying in the non-enhanced image segmentation. Moreover, the USSE introduces multi-scale attentions into the uncertainty-estimation. The multi-scale attentions increase the observational uncertainty and thus improve the estimation effectiveness to the uncertainty.

3.2 Motivations in the WSTS

3.2.1 Motivation for the DDRL

Deep Reinforcement Learning (DRL) [12] is a great candidate to transfer the tumor spatial feature from the enhanced image to the non-enhanced image (issue 1 in the teacher-student framework) via sharing detection strategies. However, traditional DRL method may cause inaccurate detection so that it is highly required to propose the new DDRL to improve the performance. 1) DRL has been applied in medical object detection [13, 14]. The detection strategy self-learned by DRL demonstrates the process to approach and locate the tumor step-by-step, it is a high-level understanding of the tumor spatial feature and thus independent of the image data distribution. Thus the teacher can transfer the spatial feature to the student by sharing the same tumor detection strategy to assist the tumor detection (Fig. 3.5). 2) With the tumor detection as a prior indicating the tumor location and size, the interference from the

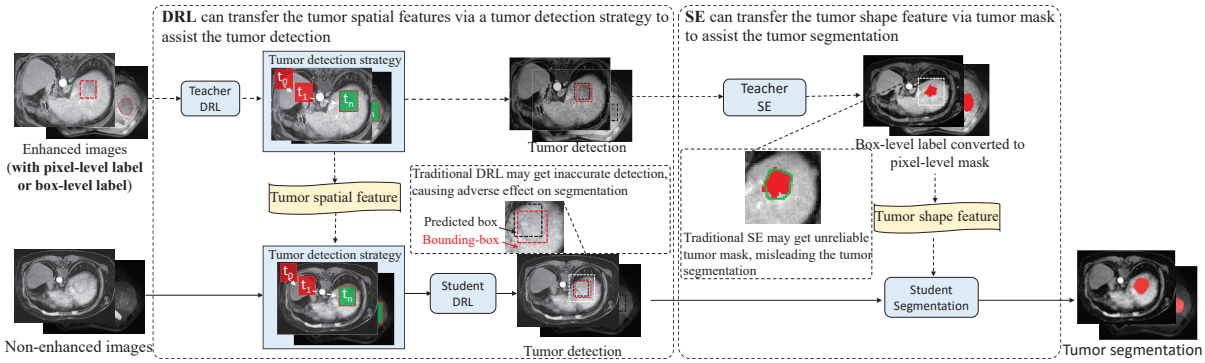


Figure 3.5. DRL is a candidate to transfer the tumor shape feature from the enhanced image to the non-enhanced image via a tumor detection strategy, thereby assisting the tumor detection in the non-enhanced image. While traditional DRL may stick into sub-optimal and cause inaccurate tumor detection, which further causes adverse effects on the subsequent tumor segmentation. SE is a candidate to transfer the tumor shape feature from the enhanced image to the non-enhanced image by predicting pixel-level tumor masks for the box-level-labeled data, thereby assisting the tumor segmentation in the non-enhanced image. While traditional SE may produce unreliable and noisy masks, which is amplified after transferred to the student and thus misleads the segmentation in the non-enhanced image.

background is removed. Thus the tumor segmentation is simplified a great extent (especially for the box-level-labeled data). 3) However, because the tumor is low-contrast in the non-enhanced image, traditional DRL may fluctuate around the optimal and thus cause inaccurate tumor detection. This inaccurate detection further causes an adverse effect on the subsequent tumor segmentation [15]. Therefore, WSTS proposes the DDRL to address this situation and promote tumor detection and segmentation performance.

3.2.2 Motivation for the USSE

Self-Ensembling (SE) [16] is a great candidate to transfer the tumor shape feature from the enhanced image to the non-enhanced image (issue 2 in the teacher-student framework) via converting the box-level label to a pseudo-pixel-level label. However, traditional SE method may mislead the tumor segmentation so that it is indispensable to propose the new USSE to promote the segmentation performance (Fig. 3.5). 1) Self-Ensembling (SE) as a semi-supervised method has been used in pixel-level segmentation with a few labeled data and large amounts of unlabeled data in medical image analysis, such as brain lesion segmentation [10] and skin lesion segmentation [17]. Thus SE has the potential to exploit the box-level-labeled data and limited pixel-level-labeled data to address the liver tumor segmentation. 2) With the tumor detection from the DDRL indicating the tumor location and the visible tumor features, SE is able to predict pixel-level tumor masks in enhanced images. For the box-level-labeled data, these

tumor masks can improve the tumor segmentation in the non-enhanced image as a pseudo label. 3) However, traditional SE may produce unreliable tumor prediction in the enhanced image considering it is weakly-supervised. The unreliable tumor prediction is amplified after transferred to the student and misleads the segmentation model training in the non-enhanced image. Therefore, WSTS proposes the USSE to improve the prediction reliability and increase the segmentation accuracy.

3.3 Related work

3.3.1 Existing work

Almost all existing work on liver tumor segmentation focuses on the enhanced image because the enhanced image presents higher contrast and clearer tumor boundary. They are grouped into the following two categories: manual methods and automatic methods.

Manual methods. Common hand-crafted feature-based methods are based on the statistical shape model [18] and its deformations, which achieves great performance in the challenge of liver disease segmentation in the 2008 MICCAI conference. Another type manual methods introduces various interactions to facilitate the tumor segmentation, such as GraphCut-based methods [19, 20], region-based threshold method [21], level set methods [22–24], B-spline transformation method [25] and machine learning methods [26–28]. Region-based segmentation methods [29, 30] mostly employ the feature similarity within the same regions for tumor segmentation. For instance, Sethi [29] proposes cancerous regions by selecting an appropriate threshold and deploys region-growing to extract connected regions according to the predefined criteria. This method makes up for the shortcomings that traditional threshold segmentation methods lack the consideration of spatial relationship among regions. Level set methods [31, 32] achieve excellent performance in tumor segmentation because they involve numerical curve and surface calculation to constrain the tumor prediction. For instance, Amarajothi [32] integrates the tumor shape and intensity as prior knowledge in level set model, thereby achieving more accurate segmentation to hepatocellular carcinoma. Machine learning methods also obtains significant performance [26–28]. For instance, Huang [26] proposed an extreme learning machine (ELM) according to random feature subspace set for liver lesion segmentation. Zhang [27] combined watershed transformation and support vector machine (SVM). Similarly, Kuo [28] combined texture feature vectors and SVM to segment liver lesions.

Automatic methods. Deep learning has attracted researchers to develop a number of automatic methods for liver tumor segmentation. These methods can be divided into three categories

according to their dimensions: 1) 2D methods, which process each individual liver image and segment the tumor, such as Cascaded-FCN [33] and multi-channel feature-fusing FCN [34]. 2) 2.5D methods, which mainly refer to exploit the spatial consistency of adjacent slices with convolutional neural network in volume images. For instance, Han [15] proposed a 2.5D FCN model, which integrated UNet with residual blocks and achieved excellent result in the ISBI LiTS challenge in 2017. 3) 3D methods, which take the whole volume image as input, such as H-DenseUNet [35].

3.3.2 Algorithm background

Deep reinforcement learning

Deep Reinforcement Learning (DRL) is a branch of machine learning, which learns from interaction and achieves great performance in medical image analysis [13, 14, 36]. DRL generally trains an agent to achieve the task goal with a Markov Decision Process: the agent observes the current state s_t of the task and decides an action a_t; with the action a_t, the task will change to the next state s_{t+1} and feedbacks the agent with a reward r. The DRL objective in such an iterative interaction process is training the agent to learn a strategy $\pi(a|s)$ that allows the agent to maximize the expected sum of reward in an episode. Eq. 3.1 shows the DRL objective, where $(s_t, a_t) \sim \rho_\pi$ represents the state that the agent observed and the action that the strategy selected at time t respectively, T represents the iteration number in the episode. Moreover, the reward is generally correlated to the task goal, namely, the agent receives the maximum reward when the agent accomplishes the task. Thus after training, the agent is able to accomplish the task goal.

$$\pi^* = \underset{\pi}{argmax} \underbrace{\sum_{t}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} r(s_t, a_t)}_{J(\pi)} \tag{3.1}$$

Practically, DRL pays more attention on future obtained rewards, so that DRL generally uses the sum of rewards with a discount γ, i.e.:

$$\pi^* = \underset{\pi}{argmax} \sum_{t}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \gamma^{T-i} r(s_t, a_t) \tag{3.2}$$

Where the discount factor $\gamma \in (0, 1)$ guarantees the sum of rewards can converge when the iteration number is infinite. And the sum of future rewards based on current state s_t is also defined as return: $G_t = r_{t+1} + \gamma r_{t+2} + \gamma^{T-t-1} r_T$.

Actor-Critic method

Actor-Critic (AC) method [37] is a type of deep reinforcement learning algorithm, which has been employed in medical image analysis [38] because of its advancement. The main workflow of the Actor-Critic method is: a strategy function as an actor selects the most probable action based on the action probabilities it estimates. A value function as a critic evaluates the value of the selected action to the task goal achieving. Then the actor modifies the action probabilities according to the evaluated value. Repeating the above procedure iteratively, the actor is able to select more optimal actions, the critic is able to evaluate the action more accurately, the Actor-Critic method is able to accomplish the task. To meet the objective of DRL, namely, learn a strategy to maximize the sum of obtained reward (Eq. 3.1), the Actor-Critic method approximates two functions with neural networks: 1) the strategy function:

$$\hat{\pi}(s,a) = P(a|s) \approx \pi(a|s) \tag{3.3}$$

Where the strategy function $\pi(s,a)$ predicts the action possibilities $P(a|s)$ when state s and outputs the action with the highest possibility. 2) the value function of strategy π (including the state-value function v or the action-value function q, because they are connected by Bellman Equation):

$$\begin{aligned} \hat{v}(s) &\approx v_\pi(s) = \mathbb{E}_\pi[G_t|s_t = s] \\ \hat{q}(s,a) &\approx q_\pi(s,a) = \mathbb{E}_\pi[G_t|s_t = s|a_t = a] \\ v_\pi(s) &= \sum \pi(a|s) q_\pi(s,a) \end{aligned} \tag{3.4}$$

Thus the optimization objective of π is to select the action with optimal action-value (or reach the state with optimal state-value) so that the agent can obtain the maximum sum of rewards, namely:

$$J(\pi) = \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \sum \gamma^{T-i} r(s_t,a_t) \tag{3.5}$$

$$= \mathbb{E}_\pi \pi(a|s)[r(s,a) + \gamma v(s_{t+1}|s_t = s)] \tag{3.6}$$

$$= \mathbb{E}_\pi \pi(a|s) q(s,a) \tag{3.7}$$

The optimization objective of state-value function and action-value function are to minimize the error between the predicted value and the true action value in Eq. 3.4. For instance, the optimization objective of action-value function can be formulated as:

$$J_q = \frac{1}{2}[q(s,a) - \hat{q}(s,a)]^2 \tag{3.8}$$

With the actor (strategy function) and the critic (value function) updating their parameters alternately, the actor selects more optimal action and the critic evaluates the action more accurately. Thus the actor-critic method is able to accomplish the task.

Experience replay

Experience replay is an important training technique in DRL, which breaks the correlations among training data (DRL exploration experience) and improves the utilization of the training data. Instead of training the DRL with the exploration experience $e_t = (s_t, a_t, r_t, s_{t+1})$ directly, experience replay stores every exploration experience into an experience memory. When updating the parameters, experience replay samples a batch of training data from the memory. Thus the correlation among inputted training data is broken, and every experience can be utilized multi-times.

3.4 Method

The WSTS leverages a Teacher Module to exploit the tumor knowledge in the enhanced image as guidance to train a Student Module, so that the Student Module is able to detect and segment the tumor from the non-enhanced image independently in the testing stage (Fig. 2.5). In detail, 1) the Teacher Module (Sec. 3.4.1) deploys the Dual-strategy DRL (DDRL) and the Uncertainty-Sifting Self-Ensembling (USSE) to obtain a tumor mask in the enhanced image. The DDRL coordinates two newly-designed Relative-entropy-biased Actor-Critics (RACs) to develop tumor detection strategies and determine the tumor location. The USSE designs a Multi-scale Uncertainty-estimation (MU) and integrates it with the Self-Ensembling (SE) to predict a pixel-level tumor mask for the box-level-label data. 2) The Student Module (Sec. 3.4.2) employs a Student DDRL (SDDRL) and a Student DenseUNet (SDUNet) to learn tumor segmentation under the guidance of the Teacher Module in the non-enhanced image. The SDDRL imitates the DDRL to learn tumor detection strategies. The SDUNet utilizes the USSE's tumor mask as a pseudo-pixel-level label (additional to the manual pixel-level label) to learn the tumor segmentation.

Advantages of the WSTS: 1) The WSTS employs the DDRL to exploit the tumor spatial feature in the enhanced image and transfer the feature to the SDDRL via the tumor detection strategy. This spatial feature transferring guides the tumor detection learning in the non-enhanced image. 2) The WSTS employs the USSE to exploit the tumor shape feature in the enhanced image and transfer the feature to the SDUNet by converting the box-level label into a pseudo-pixel-level label. This pseudo-pixel-level label (additional to the pixel-level-label) promotes

the tumor segmentation in the non-enhanced image. 3) With above two points, the WSTS retains the valuable information from the more risky modality (the enhanced image) and to assist and facilitate the segmentation in the safer one (the non-enhanced image), thereby enabling the tumor is segmented from the non-enhanced image directly.

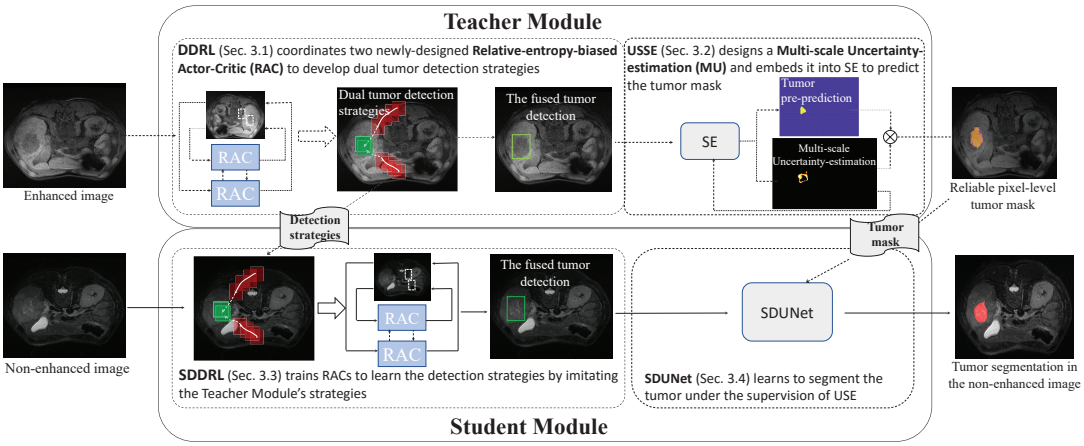


Figure 3.6. Our WSTS leverages a Teacher Module to train a Student Module so that the Student Module is able to segment the tumor from the non-enhanced image independently. The Teacher Module deploys a DDRL and a USSE to detect and segment the tumor in the enhanced image. The DDRL coordinates two newly-designed Relative-entropy-biased Actor-Critic (RAC) to develop tumor detection strategies and propose a tumor detection box. The USSE designs a Multi-scale Uncertainty-estimation (MU) and embeds it into the Self-Ensembling (SE) to predict the tumor mask in a weakly-supervised manner. The Student Module employs an SDDRL and an SDUNet to learn tumor segmentation under the guidance of the Teacher Module in the non-enhanced image. The SDDRL imitates the DDRL to learn the tumor detection strategies. The SDUNet takes the USSE's tumor mask as a pseudo-pixel-level label to learn the tumor segmentation.

3.4.1 Teacher Module

The Teacher Module employs the Dual-strategy DRL (DDRL) and the Uncertainty-Sifting Self-Ensembling (USSE) to detect the tumor in the enhanced image and predict a tumor mask for the box-level-labeled data. Thus the Teacher Student is able to guide the Student Module to detect and segment the tumor in the non-enhanced image. The DDRL cooperates two proposed Relative-entropy-biased Actor-Critics (RACs) to self-learn the tumor detection strategies and fuses their detection results to output the final optimal tumor detection. Specifically, the two RACs are required to maximize the difference between their developed strategies to avoid the inaccurate tumor detection that may be occurred in traditional DRL methods. The USSE embeds the Multi-scale Uncertainty-estimation (MU) into the SE to improve the reliability of the tumor mask. Specifically, the MU introduces multi-scale attentions into uncertainty-estimation

to increase the robustness of SE, thereby avoiding the misleading to the Student Module caused by the inaccurate tumor mask.

Advantages of the Teacher Module: 1) The Teacher Module leverages the proposed DDRL to develop the tumor detection strategy in the enhanced image. Thus the Teacher Module is able to guide the Student Module to determine the tumor location in the non-enhanced image. 2) The Teacher Module leverages the proposed USSE to convert the box-level label into a pseudo-pixel-level label (tumor mask). Thus the Teacher Module provides an accurate tumor shape feature in the box-level-labeled data for assisting the Student Module to segment the tumor.

Dual-strategy DRL (DDRL)

The DDRL (Fig. 3.7) learns to detect the tumor in the enhanced image by coordinating two newly-proposed Relative-entropy-biased Actor-Critics (RACs), thereby teaching the Student Module the image structural features to detect the tumor in the non-enhanced image. In particular, the DDRL makes the two RAC detect the same tumor with respective strategy and then fuses their detection results as the ultimate detection. The DDRL avoids the occasion in single strategy DRL that sticks into sub-optimal and leads to inaccurate tumor detection. To this purpose, the DDRL stimulates each RAC to 1) explore various strategies by maximizing the entropy of explored strategy distribution; 2) take the other RAC's decision into consideration when learning its own strategy by maximizing the relative entropy between the developed strategies. Exploring various strategies increases the exploration to the liver image and the RAC stability. Considering the other RAC's strategy prevents the two RACs from falling into the same sub-optimal and thus improves the data utilization effectiveness.

Environmental element design for the RAC. The DDRL builds respective environmental elements (state space $\boldsymbol{S}$, action space $\boldsymbol{A}$ and reward/feedback scheme $\boldsymbol{R}$) for the two RACs (*RAC* 1 & *RAC* 2) to allow them to learn respective tumor detection strategies $\pi_1(a_t|s_t)$ and $\pi_2(a_t|s_t)$. The element composition for each *RAC* i, ($i \in [1, 2]$) is designed as follows:

– *State* $\boldsymbol{s_t \in S_i}, (i \in [1, 2])$ describes the surroundings and content of the adaptive box. In the DDRL the state is a concatenation of the enhanced image and an attention map, where the attention map is built by the adaptive box (hard attention). This state arrangement allows the RAC to explore the liver image by transforming this box around the image and focus on different liver anatomy through the box. Note that, for the initial state of the *RAC* 1, the adaptive box locates in the middle of the image upper part. For the initial state of the *RAC* 2, the

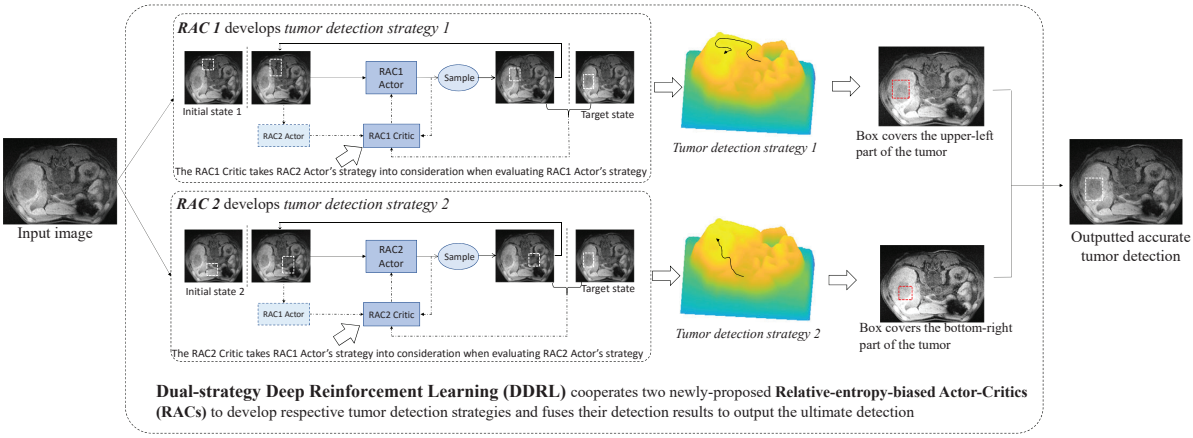


Figure 3.7. The Dual-strategy DRL (DDRL) deploys two newly-proposed RACs to detect the tumor with respective strategy and fuses their detection results as ultimate tumor detection. The DDRL avoids tumor detection sticking into sub-optimal and resulting in inaccurate detection in a single strategy DRL method. To this purpose, the DDRL stimulates each RAC to 1) explore various strategies by maximizing the entropy of explored strategy distribution; 2) take the other RAC's decision into consideration when learning its own strategy by maximizing the relative entropy between the developed strategies. Exploring various strategies increases the exploration range in the liver image and stability of the RAC. Taking the other RAC's strategy into consideration prevents the two RACs from falling into the same sub-optimal and thus improves the data utilization effectiveness.

adaptive box locates in the middle of the image bottom part. The two initial states located oppositely enable the two RACs to detect the tumor with different strategies to the greatest extent.

– *Action* $\boldsymbol{a_t \in A_i}, i \in [1, 2]$ denotes the transformations of the adaptive box in the parametric space. The action a_t is determined by the strategy $\pi_i(a_t|s_t)$ for a state s_t. In the DDRL, the action is a vector indicates the direction and stride length to translate and scale the box horizontally and vertically, and the action space $\boldsymbol{A_i}$ is continuous. Thus a_t has the ability to adjust the size and scale of the adaptive box, which allows the RAC to translate and scale the adaptive box to fit various-shaped tumors. The continuous action space means the stride length is any value in the given range instead of a few preset discrete. The continuous action space allows the RAC to select the stride length flexibly according to the relative distance between the adaptive box and the tumor bounding-box.

– *reward function* $\boldsymbol{r \in R_i}$ encodes the feedback from the label to the action for the agent network training. In the DDRL, both RACs have the same reward function formula. The reward function formula denotes the reward as the exponent of the detection accuracy improvement

caused by the action:

$$r(s_t, a_t) = \begin{cases} e^{-d_t} & \Delta > 0 \\ 0 & \Delta \leq 0 \end{cases} \tag{3.9}$$

Where d_t is the detection accuracy after the action a_t, which is evaluated by the sum of the Euclidean distance between the adaptive box's four vertices and the bounding-box's four vertices. Δ is the decreasing of d caused by the action, namely, $\Delta = d_{t-1} - d_t$. When the adaptive box coincides with the tumor bounding box, the reward is maximum (i.e., $r = 1$). Thus, to obtain the maximum sum of rewards, each RAC learns to 1) transform the adaptive box with optimal action to approach the tumor bounding-box, so that $\Delta > 0$ and $r > 0$; 2) transform the adaptive box to match the tumor bounding-box, so that $d_t = 0$ and $r = 1$.

Learning objective for the RAC. The DDRL designs following optimization objective for each RAC to address the tumor detection (here taking *RAC* 1 as an example, the optimization of *RAC* 2 follows the same principle):

$$J(\pi_1) = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_{\pi_1}} \{r(s_t, a_t) + \alpha[\mathcal{H}(\pi_1(\cdot|s_t)) + \mathcal{D}_{KL_1}(\pi_1(\cdot|s_t)\|\pi_2(\cdot|s_t))]\} \tag{3.10}$$

Where $\mathbb{E}_{(s_t,a_t)\sim\rho_{\pi_1}} r(s_t, a_t)$ is the objective of traditional DRL, namely, the sum of obtained rewards. Maximizing this item aims to guide the *RAC* 1 to approach and achieve the task goal. $\mathcal{H}(\pi_1(\cdot|s_t))$ represents the entropy of the strategy, which is inspired by the Soft Actor-Critic (SAC) method [39]. Maximizing this entropy aims to guide the *RAC* 1 to approach the target goal with different strategies. In other words, it guides the *RAC* 1 to explore the liver image as random as possible, as well as improves the *RAC* 1 stability. $\mathcal{D}_{KL_1}(\pi_1(\cdot|s_t)\|\pi_2(\cdot|s_t))$ denotes the relative entropy between RACs. Maximizing this item aims to stimulate the *RAC* 1 to develop a distinct strategy from the *RAC* 2 developed. Maximizing the relative entropy prevents two RACs from falling into the same strategy, thereby improving the tumor detection performance. In particular, when optimizing the *RAC* 1, the DDRL takes the decision made by the *RAC* 2 into consideration and enables the *RAC* 1 to make a different decision from the *RAC* 2 made. α is a temperature parameter that determines the relative importance of the entropy against the reward. For the following derivation, we will omit writing the temperature explicitly, as it can always be subsumed into the reward by scaling it by α^{-1}.

Optimization procedure for the learning objective. To make each RAC achieve the above learning objective, the DDRL refers to the actor-critic method. In particular, the DDRL constructs a new state-value function V_1 and a new action-value function Q_1 (as the critic) to

evaluate and guide a strategy function π_1 (as the actor) to develop the strategy. The state value function V_1 for the strategy π_1 is trained to minimize the squared residual error:

$$J(V_1) = \mathbb{E}_{s_t \sim \mathcal{D}} \left\{ \frac{1}{2} (V_1(s_t) - \mathbb{E}[Q_{a_t \sim \pi_1}(s_t, a_t) - log\pi_2(a_t|s_t)_{a_t \sim \pi_2}])^2 \right\} \tag{3.11}$$

where $\mathcal{D}$ is a replay buffer sampled from the replay-experience memory (refer to Sec. 3.3.2). $\pi_2(a_t|s_t)$ acts as a bias term, which takes the action a_t made by π_2 under the state s_t into consideration. The derivation from Eq. 3.10 to Eq. 3.11 can be seen in E. Based on the Eq. 3.11, the value functions of two RACs are optimized alternately. Namely, when updating one RAC, the other RAC only provides a bias term instead of participating in updating. The action-value function Q_1 can be trained to minimize the Bellman residual:

$$J(Q_1) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}} [\frac{1}{2} (Q_1(s_t, a_t) - \hat{Q}_1(s_t, a_t))^2] \tag{3.12}$$

with

$$\hat{Q}_1(s_t, a_t) = r(s_t, a_t) + \zeta \mathbb{E}_{s_{t+1}}[V_1(s_{t+1})] \tag{3.13}$$

The DDRL represents the strategy function π_1 with the distribution of the action-value function Q_1 (refer to Eq. 3.5), namely:

$$\pi_1(s_t, a_t) \propto exp(Q_1(s_t, a_t)) \tag{3.14}$$

Thus the updating of strategy function π_1 is to make the distribution of strategy closer to the distribution of Q_1. Here, refer to SAC[39], the DDRL employs and minimizes the Kullback-Leibler divergence of the distributions, i.e., the DDRL aims to find the strategy when the KL divergence of the strategy distribution and action-value distribution is minimum:

$$J(\pi_1) = \mathbb{E}_{s_t \sim \mathcal{D}} D_{KL_2}(\pi_1(\cdot|s_t) \| \frac{exp(Q_1(s_t, \cdot))}{Z(s_t)}) \tag{3.15}$$

where the partition function $Z(s_t)$ normalizes the distribution, it does not contribute to the gradient with respect to the new strategy and can thus be ignored. The networks to approximate the state-value function, action-value function, and the strategy function can be seen in F.

Fusing the tumor detection from the dual RACs. After the dual RACs develop respective tumor detection strategies and propose respective tumor detection boxes, the DDRL adopts the union of two boxes as the ultimate detection result for subsequent tumor segmentation. Compared with one detection box from a single strategy that usually covers most parts of the tumor area instead of the complete tumor, the union of two detection boxes from two totally various

strategies covers more tumor areas. Thus the DDRL improves the detection quality effectively and outputs a more accurate tumor detection box for subsequent tumor segmentation.

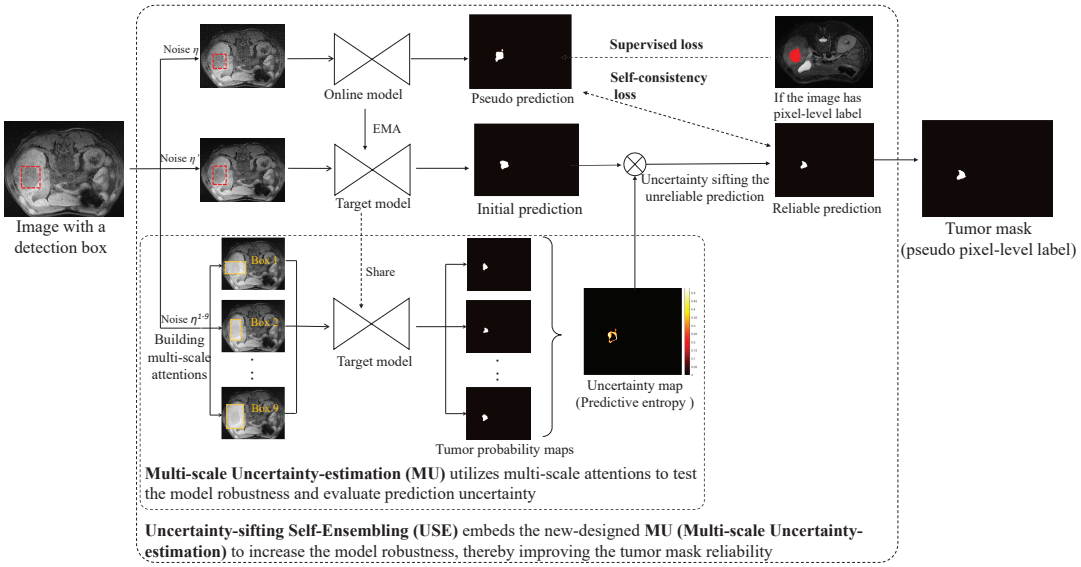


Figure 3.8. The Uncertainty-Sifting Self-Ensembling (USSE) builds an online model and a target model to build the self-consistency loss and exploit the box-level-labeled data. The prediction of the target model sifted by the Multi-scale Uncertainty-estimation (MU) is employed as a reliable prediction to constrain the segmentation of the online model. The MU creatively introduces multi-scale attentions into the uncertainty-estimation, which increases the observational uncertainty and thus improves the estimation effectiveness of the uncertainty to the tumor prediction.

Uncertainty-Sifting Self-Ensembling (USSE)

The USSE (Fig. 3.8) predicts tumor masks for the box-level-labeled data by integrating the Self-Ensembling (SE) with a proposed Multi-scale Uncertainty-estimation (MU), thereby converting the box-level label into a pseudo-pixel-level label (additional to the manual pixel-level label) to improve the Student Module. SE as a weakly-supervised method segments the tumor with partial pixel-level-labeled data and partial box-level-labeled data. MU as a newly proposed uncertainty-estimation method sifts the reliable segmentation from the unreliable and improves the segmentation accuracy. The USSE consists of two models (an online model and a target model). The target model's parameter is the ensemble of the online model's. Both models take the enhanced image and the DDRL detection box as input, and output tumor segmentation. Considering the ensemble segmentation from the target model is more accurate than that from the online model [10], the online model learns from the target model by minimizing the consistency between the two segmentation. However, without pixel-level ground-truth, the

predicted segmentation may be unreliable and noisy. For this point, the USSE designs MU to sift the reliable segmentation from the unreliable. Therefore, the training loss for the USSE is formulated as:

$$\mathcal{L} = \overbrace{\mathcal{L}_s[f(x,\eta,\delta),y]}^{\text{supervised loss}} + \gamma \overbrace{\mathcal{L}_g[f(x,\eta,\delta),f(x,\eta',\delta')]}^{\text{self-consistency loss}} \tag{3.16}$$

The first item in the above equation right part is the supervised loss. In particular, the USSE builds the online model and the target model with the same Dense U-Net architecture [40] and respective parameters δ and δ'. The δ' is updated by an exponential moving average (EMA) strategy every time after the parameters δ is updated (namely, $\delta' = \beta\delta' + (1-\beta)\delta$, where β is the EMA decay to control the updating rate). For an enhanced liver image with pixel-level label y, the USSE concatenates the image and detection box as the input x, adds random noise η, and feeds the input into the online model. Then according to the supervised loss $\mathcal{L}_s[f(x,\eta,\delta),y]$ to update the online model's parameters. Here, the supervised loss $\mathcal{L}_s$ is Dice loss which performs greatly in imbalanced-task (the imbalance between the tumor area and the background area). γ is a ramp-up weighting coefficient [41] that controls the trade-off between the two items in the above equation.

The second item in the above equation right part is self-consistency loss. No matter the inputted data has a pixel-level label, the target model's prediction $f(x,\eta',\delta')$ is deployed as a potential target to restrict the online model's prediction $f(x,\eta,\delta)$. Namely, the online model is also updated by the self-consistency loss $\mathcal{L}_g[f(x,\eta,\delta),f(x,\eta',\delta')]$:

$$\mathcal{L}_g[f(x,\eta,\theta),f(x,\eta',\theta')] = \frac{\sum \Pi(u<H)\|f(x,\eta,\theta),f(x,\eta',\theta')\|_2}{\sum \Pi(u<H)} \tag{3.17}$$

Where $\|\cdot\|_2$ is the pixel-to-pixel square distance to measure the self-consistency loss between two predictions. More importantly, to increase the reliability of the self-consistency loss, an indicator function $\Pi(u < H)$ is deployed to filter the high-uncertainty prediction. u is the Multi-scale Uncertainty-estimation (MU) to the prediction of every pixel and H is an adaptive uncertainty threshold. In particular, considering different attentions lead to different segmentation [42], MU introduces multi-scale attentions into the uncertainty-estimation to increase the observation uncertainty. As shown in Fig. 3.8, MU adapts the original detection box to nine new attention boxes and builds nine different inputs ($\{x_t\}_{t=1}^{T}, T = 9$). These nine boxes contain three sizes (original size, original size *1.2, original size * 1.4), and three scales (1:1, 1:2, 2:1) for each size. Then, MU introduces Gauss noise to these nine inputs ($\{x_t\}_{t=1}^{T}, T = 9$) and feeds them into the target model during forward propagation. These nine different inputs result in nine tumor probability maps ($\{P_t\}_{t=1}^{T}, T = 9$). Then the uncertainty is evaluated according to

the predictive entropy u and the average prediction μ (Eq. 3.18).

$$\begin{aligned} \mu &= \frac{1}{T}\sum_{t=1}^{T} P_t \\ u &= -\mu log\mu \end{aligned} \tag{3.18}$$

For each input, H is determined adaptively by the overall uncertainty of this input, it can be formulated as: $H = \beta' u_{max} + (1 - \beta')u_{min}$. β' is a time-dependent warming up function [16, 43] to increase the uncertainty threshold gradually.

3.4.2 Student module

The Student Module consists of a Student DDRL (SDDRL) and a Student Dense U-Net (SDUNet) to receive the guidance and assistance from the Teacher Module to learn to detect and segment the tumor in the non-enhanced image. The SDDRL imitates the DDRL to learn the tumor detection strategies and determines the tumor location in the non-enhanced image for subsequent segmentation. The SDUNet learns to segment the detected tumor under the supervision of the pseudo-pixel-level label from the USSE and the manual pixel-level label. After the above learning process, in the testing stage, the Student Module is deployed to detect and segment the tumor from the non-enhanced image independently without the assistance of the Teacher Module.

Advantages of the Student Module: 1) The Student Module learns the tumor spatial feature in the non-enhanced image under the guidance of the Teacher Module instead of self-exploring. This guidance prevents the Student Module from divergence because the tumor feature is weak and accelerates the detection strategies learning. 2) The Student Module learns to extract the tumor shape feature under the guidance of the pseudo-pixel-level label instead of the supervision from the box-level label. This pseudo-pixel-level label enables the Student Module to focus on the correct tumor area and avoid the interference from surrounding tissues.

Student Dual-strategy DRL (SDDRL)

The SDDRL is built based on the DDRL architecture to learn to detect the tumor in the non-enhanced image. During the training stage, the SDDRL follows the DDRL's strategies to transform the adaptive box iteratively and determine the tumor location and size in the non-enhanced image. Such iterative transforming allows the SDDRL learns to select optimal actions to approach the tumor when it observes different liver anatomy through the adaptive box. Thus during the testing stage, the SDDRL is able to select actions step by step to detect the tumor independently in the non-enhanced image. More importantly, because the weak features

in the non-enhanced image usually cause repetitive and useless explorations. By following the DDRL's actions step by step to learn the tumor detection strategies instead of self-exploring the non-enhanced image, the SDDRL is able to converge fast and avoids over-fitting.

For the framework and the environmental element design, the SDDRL builds the same dual RACs as the DDRL and the same element compositions (the state, the action, and the reward function). In detail, 1) the SDDRL concatenates the non-enhanced image and the adaptive boxes when constructing the states for the two RACs, note that the initial adaptive boxes here are the same as the initial boxes in the DDRL; 2) the RACs follows the DDRL's every action that transforms the adaptive boxes step by step to explore the non-enhanced image; 3) the RACs obtains the same rewards as the DDRL after each action, because the same box and the same action lead to the same tumor detection accuracy.

For the learning optimization objective and optimization procedure, the SDDRL learns to develop the strategies from the DDRL in the non-enhanced image. In particular, the SDDRL deploys a critic to estimate the state-value function and the action-value function, as well as an actor to estimate the strategy function based on the observed information (state) and the knowledge (action, and reward) from the DDRL. In our implementation, these three functions are built and trained with the same process as the DDRL. However, the above critic and actor in the SDDRL can be trained according to the principle of a traditional actor-critic method (Sec. 3.3.2), because the DDRL has provided desirable strategies for the SDDRL.

For the fusing of detection results, the SDDRL follows the DDRL and adopts the union of two detection boxes as the ultimate detection result for further tumor segmentation. Because in the non-enhanced image the tumor is low-contrast, the tumor detection box predicted by a single detection strategy usually has a little deviation compared with the optimal bounding-box. This deviation causes that the tumor area outside the box is missed in segmentation since the segmentation model focuses attention on inside the box, thereby reducing the segmentation accuracy. Adopting the union of two detection boxes effectively expands the attention area, so that the tumor segmentation is complete and high accuracy.

Student Dense U-Net (SDUNet)

The SDUNet learns to segment the tumor from the non-enhanced image based on the attention from the SDDRL and the supervision from the USSE. The SDUNet builds a hard attention map according to the tumor detection box proposed by the SDDRL, and concatenates the attention map with the non-enhanced image together as the input. The SDUNet deploys a Dense U-Net

[40] as the main segmentation model. The SDUNet learns to segment the tumor under the supervision of the pseudo-pixel-level label from the USSE (in the box-level-labeled data) and the manual pixel-level label (in the pixel-level-labeled data). The training loss for the SDUNet is:

$$\mathcal{L} = \gamma' \overbrace{\mathcal{L}_b[F(x,\eta), f(x,\eta)]}^{\text{box level label}} + \overbrace{\mathcal{L}_p[F(x,\eta), y]}^{\text{pixel level label}} \tag{3.19}$$

Where γ' is the ramp-up weighting coefficient [41] that controls the trade-off between the supportive ground-truth (box-level-labeled data) and the manual ground-truth (pixel-level-labeled data). $\mathcal{L}_b$ and $\mathcal{L}_p$ are Dice loss between the tumor segmentation ($F(x,\eta)$) and the ground-truth ($f(x,\eta)$ or y). Dice loss effectively addresses the imbalance between the area of background and the area of the tumor [44]. Moreover, the Dice coefficient is one of the most important matrices to evaluate the tumor segmentation performance, thus taking Dice as the loss function promotes the model to improve segmentation accuracy.

With the tumor detection box from the SDDRL as hard attention, the SDUNet is able to focus on the tumor region instead of being interfered by the background. Such attention-focusing is crucial for the small and low-contrast tumor segmentation. With the pseudo-pixel-level from the USSE and manual label as the supervision, the SDUNet is able to learn to extract the tumor inconspicuous feature in the pixel-level from the non-enhanced image systemically. With the Dense UNet that integrates two effective tools (the U-Net [45] and Dense Block [46]) as the segmentation network, the SDUNet retains the information flow and gradient flow in the model maximally to segment the tumor details.

3.5 Experiment

3.5.1 Data acquirement

In the validation, a total of 250 patients with liver axial MRI images are selected, which include 150 Hemangioma, 100 Hepatocellular Carcinoma. All subjects obeyed initial standard clinical liver MRI protocol examinations. The non-enhanced liver MRI images (T2WI) were obtained by a 3-T MRI system (GE Signa), where the image size is 256×256, the slice thickness is 6mm. The enhanced liver MRI images (T1FS) using Fat Saturation sequence 20 seconds-10 minutes after intravenous injection of a gadolinium-based contrast agent (0.1 mmol/kg; Bayer Schering Pharma AG, Berlin, Germany), where the image size is 512×512, the slice thickness is 4mm. A radiologist with 7 years of experience in MR Liver imaging analyzed the enhanced images and labeled manual contour of the dominant index lesion in each scan as ground truth. WSTS

randomly selects 200 patients for training, and uses the remaining 50 patients for independent testing.

3.5.2 Implementation details

Input preprocessing. In the training stage, all the original non-enhanced T2WI images of 256×256 are fed into WSTS. The slice of the enhanced T1FS image and the ground-truth that correspond to the non-enhanced image are selected and resized to 256×256. The box-level label and the pixel-level label for tumor detection and tumor segmentation are deduced from the manual contour. Namely, regarding the bounding-box of manual contour as the box-level label and highlighting the inside of manual contour as the pixel-level label. In the testing stage, only the non-enhanced T2 WI image is employed.

Hyper-parameter setting. The experience replay memory sizes for each Teacher RAC and Student RAC are 20000 and 30000 respectively. For the learning rate, the initial rate in the Teacher Module and Student Module starts with $1e^{-3}$ and $1e^{-4}$ respectively, ends with $1e^{-6}$ based on the decay coefficient 0.8. The replay buffer size in all DRL methods is 80. The iteration step in the DDRL and the SDDRL for each episode is 35. For the optimizer, all DRL methods deploy Adam, other methods deploy SGD. The initial box size when building the state in the DDRL and the SDDRL is $24 * 24$. The step sizes of translation and scale in the action are $[0-6]$ pixels and $[0-4]$ pixels respectively. ζ in the DDRL is 0.99. The temperature parameter in the DDRL α is set as 1.0. β in the EMA is 0.99 to update the target model parameters. The range of Gauss noise in the USSE is [-0.1, 0.1]. Drop-out is deployed with rate of 0.1 in every convolution layer and fully connected layer.

Software and hardware setting. WSTS is implemented with the PyTorch deep learning framework (version 1.4.1) with 8 NVIDIA GTX 1080 GPUs and 48 Intel Xeon CPUs under Ubuntu 16.04. The CUDA version is 10.0 and the GPU driver version is 410.78.

3.5.3 Evaluation criteria

Our experiments evaluate tumor segmentation results with the following criteria. The tumor possibility threshold for every pixel is 0.75, namely, when the prediction possibility for a pixel is higher than 0.75, this pixel is classified as a tumor pixel, otherwise it is classified as a background pixel.

Recall measures the rate that the prediction is correct out of the tumor segmentation. Recall is

formulated as:

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{3.20}$$

where TP is the number of pixels that are predicted as the tumor and consistent with the ground-truth. FN is the number of pixels that are predicted as background but inconsistent with the ground-truth, namely the pixels actually belong to tumors. Recall value locates between 0 to 1. The value is closer to 1 means more tumor pixels are classified correctly.

Dice coefficient evaluates the overlap between the tumor prediction and the ground-truth. Dice is formulated as:

$$Dice = \frac{2 \times |P \cap G|}{|P \cup G|} \times 100\% \tag{3.21}$$

where $|P \cap G|$ counts the number of pixels that belong to the tumor prediction P and ground-truth G. $|P \cup G|$ counts the number of pixels that belong to the tumor prediction P or belong to the ground-truth G. Dice value locates between 0 to 1. The value is closer to 1 indicates more tumor pixels in the prediction are correct and fewer tumor pixels in the ground-truth are missed.

95% Hausdorff distance (95HD) evaluates the respective distance between the tumor boundary in the tumor prediction and the tumor boundary in the ground-truth. Firstly, Maximum HD is defined as:

$$HD(p, g) = max(h(p, g), h(g, p)) \tag{3.22}$$

where p, g denote the tumor boundary point set respectively in the prediction P and the ground-truth G. While $h(g, r)$ is called the directed Hausdorff distance and is given by $h(g, r) = max_{g_i \in g} min_{r_i \in r} \|g_i - r_i\|$, where $\|g_i - r_i\|$ is Euclidean distance in our evaluation. 95HD is based on the calculation of the 95th percentile of the maximum HD between boundary points in the prediction and the ground-truth with the purpose of eliminating the impact of a very small subset of the outliers. 95HD value is smaller indicates the prediction boundary is more consistent with the ground-truth boundary, which means the tumor prediction is more accurate.

Cohen Kappa Coefficient (KAP) [47] is a measure of agreement between the segmentation and the ground-truth. As an advantage over other evaluation matrices, KAP is more robust because it takes into account the agreement caused by chance. KAP is given by:

$$KAP = \frac{f_a - f_c}{N - f_c} \tag{3.23}$$

Where N is the total number of observations, in our case it is the pixel numbers. f_a and f_c can

be expressed in terms of four overlap cardinalities:

$$\begin{aligned} f_a &= TP + TN \\ f_c &= \frac{(TN+FN)(TN+FP)+(FP+TP)(FN+TP)}{N} \end{aligned} \tag{3.24}$$

Where TP and FN have been defined in the Recall. TN is the number of pixels which are predicted as background correctly. FP is the number of pixels which are predicted as tumor but actually belong to the background.

Receiver Operating Characteristic (ROC) curve is drawn based on the confusion matrix (Specificity and Sensitivity), which are usually used to evaluate the classification ability of deep learning methods to every pixel in a given dataset. ROC curve reflects the relationship between the sensitivity and specificity, which is a comprehensive representation of segmentation accuracy. ROC curve is more convex or closer to the upper left corner, indicating that the evaluation value of the ROC curve is greater. The Area Under Curve (AUC) can represent the above ROC evaluation value quantitatively. ROC curves simply and intuitively illustrate the segmentation performance, thus the clinical accuracy of segmentation method can be judged by naked eyes directly.

3.5.4 Experimental setting

Control experiments

Considering that deploying *a normal DRL and a normal SE* to segment the tumor from non-enhanced images directly as the baseline method, control experiments evaluate the efficiency of newly-proposed innovations to the liver tumor segmentation. Particularly, our control experiments 1) implement the baseline method and evaluate the segmentation result with the above metrics; 2) employ the weakly-supervised teacher-student framework (TCH-ST) on the basis of the baseline method, namely, exploit the enhanced image to assist during training, and evaluate the result with the above metrics. Specifically, both Teacher Module and Student Module take a normal DRL to learn the tumor detection strategy, and the Teacher Module utilizes a normal SE to assist the Student Module; 3) replace the normal DRL in the above TCH-ST with the newly-proposed DDRL and evaluate the segmentation result with the above criteria to investigate the effectiveness of the DDRL; 4) replace the normal SE in the above TCH-ST with the newly-proposed USSE and evaluate the segmentation with the above criteria to investigate the advantage of the USSE.

Ablation experiments

Based on the WSTS framework, the ablation experiment firstly compares the influence caused by different DRL methods to the ultimate segmentation. In particular, the experiment replaces SAC in DDRL with Deep Q-learning (DQN [12]) and Advantage Actor-Critic (A2C) respectively to display the advantage of SAC in this task. DQN is the most typical DRL method deployed in medical image analysis ([13][36]). A2C is a continuous action space DRL method, which is also employed in recent years[38]. The ablation experiment employs two DQNs and two A2Cs to achieve the dual tumor detection strategies, where the relative-entropy maximization is also deployed as the same as the DDRL.

Based on the WSTS framework, the ablation experiment also compares the effectiveness caused by different semi-supervised methods to the ultimate segmentation, namely, compares the ability of different methods to exploit the box-level-labeled data. In particular, the experiment deploys DAN [48], TCSE[17] and, MCSE [41] to replace the USSE. The DAN adopts two adversarial networks to enable the tumor prediction in the non-pixel-level-labeled data to approach the correct segmentation. The TCSE utilizes segmentation consistency before and after the image transformation. The MCSE embeds the Monte-Carlo uncertainty-estimation into a normal SE to output a reliable prediction.

Inter-comparison experiments

Inter-comparison experiments compares our method with other two kinds baseline methods (U-Net[45], Mask R-CNN[49]), and two state-of-art methods (RgGAN [11] & PSCGAN [50]). U-Net and Mask R-CNN are used widely in medical image segmentation, they are trained and tested only on the non-enhanced image. RgGAN and PSCGAN are assisted by the enhanced image during the training stage. More particularly, PSCGAN is an advanced method to segment the ischemic heart image, which synthesizes the enhanced image firstly and then inputs both non-enhanced and synthesized enhanced images to get the segmentation. The inter-comparison experiments train the above methods with all pixel-level-labeled data (200 patients) and half pixel-level-labeled data (100 patients) respectively, and evaluate the validation result with the criteria.

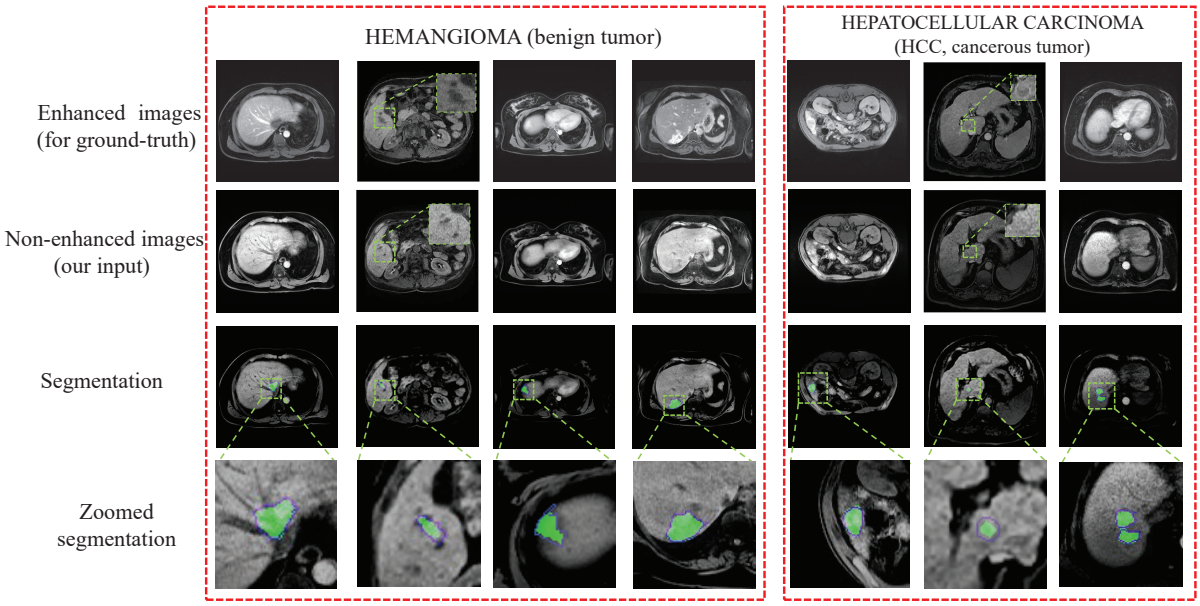


Figure 3.9. Tumor segmentation visualization presents that WSTS achieves a high overlap between the tumor prediction (green area) and the ground-truth (blue contour). In the non-enhanced image, the tumor is low-contrast and barely visible compared with that in the enhanced image. The small size of some tumors aggravates the difficulty to identify and segment the tumor from the non-enhanced image. However, the visualization shows that WSTS has the ability to identify and segment the tumor from the complex background effectively and accurately.

3.5.5 Experimental result

Comprehensive analysis

Qualitative evaluation of WSTS performance. Fig. 3.9 displays some segmentation samples visually compared against the ground-truth. The high overlap between the segmentation (green mask) and the ground-truth (blue contour) demonstrates that WSTS effectively segments tumors from non-enhanced images with excellent accuracy. The displayed images from two kinds of tumors (Hemangioma and HCC) cover various tumor contrasts, including clearly visible, barely visible, and invisible to naked eyes. WSTS demonstrates great robustness by overcoming the low-contrast of these tumors. In Fig. 3.9, every tumor is identified accurately, which indicates that WSTS has the ability to develop the tumor detection strategy and determine the tumor location in the non-enhanced image. The high coincidence between the mask contour and the ground-truth contour indicates that WSTS is capable to capture tumor features and segment the tumor accurately.

Quantitative evaluation of WSTS result. Based on the quantitative result in the last row

Table 3.1. Control experimental result.

Method	label used		Metrics			
	pixel-level	box-level	Dice[%]	95HD[mm]	Recall[%]	KAP [%]
Baseline	100	100	51.65±24.66	4.83±1.14	56.16±30.20	51.58±24.67
TCH-ST	100	100	74.93±19.93	4.28±1.06	82.48±22.67	74.86±19.94
TCH-ST+DDRL	100	100	79.68±18.44	4.16±1.07	84.69±15.71	79.64±18.45
TCH-ST+USSE	100	100	79.71±11.41	4.11±0.94	84.39±11.52	79.66±10.44
WSTS	100	100	**83.11±12.11**	**3.96±0.96**	**85.12±12.46**	**83.07±11.12**

of Tab. 2.1, WSTS overall obtains accurate tumor segmentation. In detail, WSTS achieves 83.11±12.11% in Dice, 3.96 ± 0.96 mm in 95HD, 85.12 ± 12.46% in Recall, and 83.07 ± 11.12 in KAP. The high Dice and Recall values indicate the SDDRL detects the tumor accurately, and the SDUNet distinguishes most tumor pixels from the non-enhanced image correctly. The low 95HD value indicates that even in the weakly-supervised manner, the USSE facilitates WSTS to predict the tumor boundary accurately. The high KAP value indicates the SDUNet has a strong classification ability to classify every tumor pixel instead of predicting by chance. It also illustrates the pseudo label from the USSE is effective. The ROC curve in the Fig. 3.10 illustrates the strong tumor segmentation ability of WSTS. The AUC of the ROC is close to 1.0, which proves from another aspect that WSTS has a great ability to classify each pixel correctly.

Evaluation of control experiments

Based on the evaluation result in Tab. 2.1 and Fig. 3.10, the control experiment validates the DDRL and the USSE improve the tumor segmentation accuracy and stability effectively.

The baseline method performs poorly to segment the tumor from the non-enhanced image directly. According to Tab. 2.1, the baseline method achieves 51.65 ± 24.66% in Dice, 4.83 ± 1.14 mm in 95HD, 56.16 ± 30.20% in Recall, and 51.58 ± 24.67 in KAP. The low mean Dice and Recall values, high mean 95HD value and low KAP value indicate the baseline method has no capability to segment the tumor from the non-enhanced image. In other words, almost half of tumor pixels are classified falsely, and the segmentation contour coincides with the ground-truth poorly. The gentle ROC curve in Fig. 3.10 also explains the segmentation ability of the baseline method is weak, where the AUC (area under the curve) is only almost 0.6.

The weakly-supervised teacher-student framework (TCH-ST) addresses the tumor segmentation from the non-enhanced image by exploiting the enhanced image and assisting the segmentation training in the non-enhanced image. Based on the Tab. 2.1 and Fig. 3.10, after the TCH-ST is employed, the segmentation performance is greatly improved from all aspects.

Dice value reaches 74.93±24.66%, 95HD decreases to 4.28 ± 1.06 mm, Recall value and KAP value increases to 82.48 ± 22.67% and 74.86 ± 19.94% respectively, and the ROC curve also becomes steep. The Dice, Recall and KAP improvements indicate that the TCH-ST is more effective to distinguish tumor pixels correctly from the background. The decreased 95HD indicates TCH-ST has a stronger capability to capture the tumor shape features by employing SE in the enhanced image compared with employing SE directly in the non-enhanced image. The standard deviation decrease for the above matrices also indicates the segmentation stability increases.

The DDRL improves tumor segmentation accuracy by increasing the quality of tumor detection. After the DDRL is deployed in the weakly-supervised teacher-student framework (TCH-ST), compared with the normal DRL, the mean Dice value increases by 4.75%, the mean 95HD value decreases 0.08 mm, the mean Recall value increases 2.21%, and the mean KAP value increases 4.78%. The ROC curve also becomes steeper. Such segmentation improvement indicates the DDRL proposes a more accurate tumor detection box than the normal DRL. The dual RACs in the DDRL explores larger image areas and proposes the detection box that is closer to the tumor bounding-box. The accurate tumor detection enables the USSE (and the SDUNet) to focus attention on the optimal tumor region and get more accurate segmentation.

The USSE improves the segmentation performance in the tumor edge area and the segmentation stability by increasing the tumor prediction reliability. After the USSE is deployed in the weakly-supervised teacher-student framework (TCH-ST), compared with the normal SE, the Dice value increases by 4.78%, the 95HD value decreases 0.17 mm, the Recall value increases 1.91%, and the KAP value increases 4.80%. The ROC curve also closer to the upper-left corner. Compared with the DDRL, the USSE mainly improves the Dice and 95HD, which indicates the Multi-scale Uncertainty-estimation (MU) in the USSE increases the tumor prediction reliability in the tumor edge area. The standard deviation also declines more compared with the TCH-ST and the DDRL, which demonstrates MU promotes model stability and robustness.

Evaluation of ablation experiments

According to Fig. 3.11, the SAC method that WSTS employed is the best to facilitate tumor segmentation. In the ablation experiments, after DQN is deployed in the DDRL to learn the detection strategy, the experiment result achieves 67.09% in Dice, 4.32 mm in 95HD, 71.00% in Recall, and 68.97% in KAP, where Dice, Recall, and KAP decrease 16.02%, 14.12%, and 14.10% respectively, and 95HD increases 0.36 mm compared with the SAC. The reason for such a poor result may be that the DQN is a discrete action-space method, its detection box

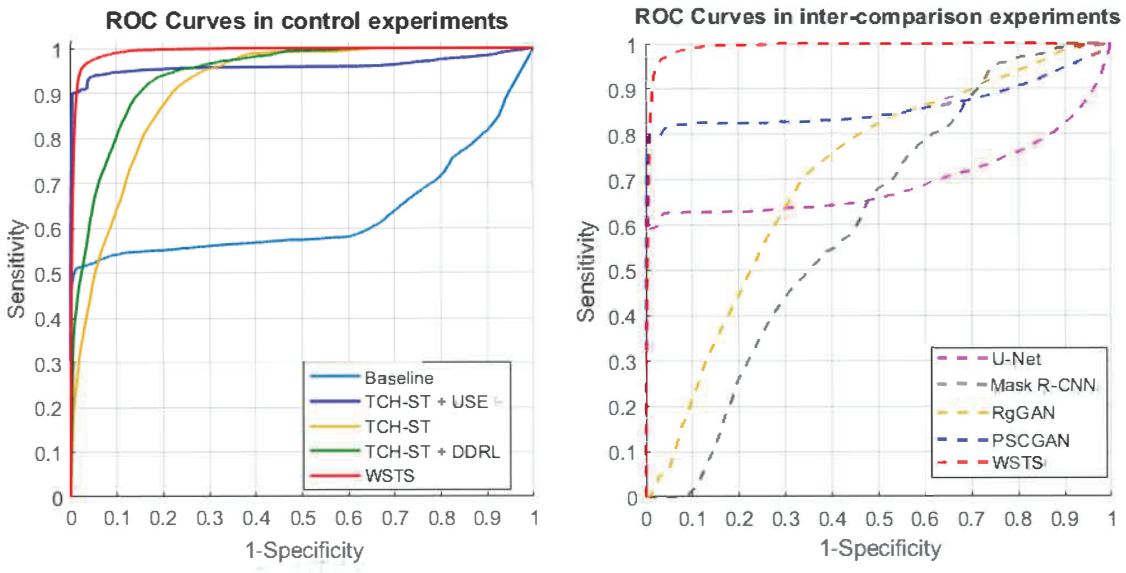


Figure 3.10. The ROC Curves of control experiments (left) and inter-comparison experiments (right) show every part of WSTS increases AUC to varying degrees and WSTS has the largest AUC compared with other fully-supervised methods. In the left figure, compared with the baseline method (without assisting of the enhanced image), the teacher-student framework (TCH-ST), the DDRL, and the USSE improve the model segmentation performance respectively. In the right figure, with the same amount of pix-level-labeled data, our WSTS has the largest AUC compared with other fully-supervised methods by leveraging the additional box-level-labeled data.

does not focus the segmentation attention on the tumor optimally and thus causes interference to the tumor segmentation. After the A2C is deployed in the DDRL, the segmentation achieves 78.31% in Dice, 4.04 mm in 95HD, 82.91% in Recall, and 78.22% in KAP, where Dice, Recall, and KAP decrease 4.80%, 3.21%, and 4.85% respectively, and 95HD increases 0.08 mm compared with SAC. Such accuracy decreasing may be caused by the delayed-reward in the A2C method, which results in the detection box is less accurate than the box proposed by the SAC.

According to Fig. 3.12, the proposed USSE is most capable to exploit the box-level labeled data and improve the segmentation performance compared with the other three state-of-art methods. Under the same amount of pixel-level-labeled data and the same tumor box provided by the DDRL, the TCSE achieves 78.34% in Dice, 4.19 mm in 95HD, 80.80% in Recall, and 78.29% in KAP. The DAN achieves 68.77% in Dice, 4.35mm in 95HD, 73.74% in Recall, and 68.67% in KAP. The MCSE achieves 80.30% in Dice, 4.16 mm in 95HD, 84.68% in Recall, and 80.25% in KAP. Based on the above results, SE-based methods (TCSE, MCSE, and USSE) compared with DAN performs better in the tumor segmentation task, where the Dice value

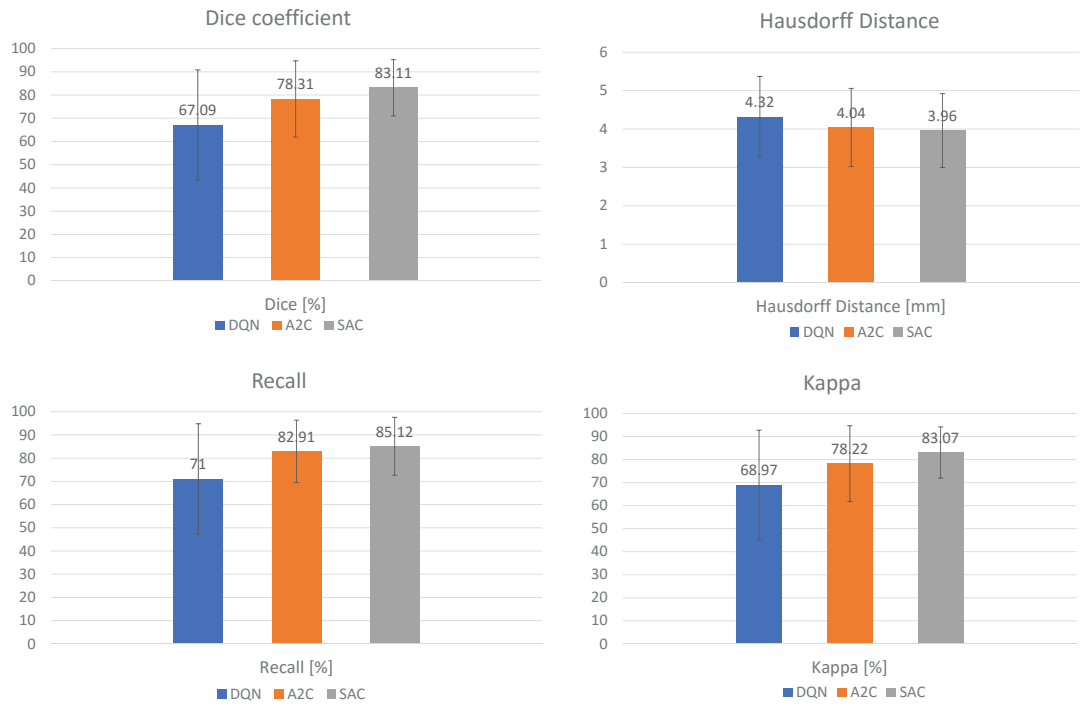


Figure 3.11. Ablation experimental result related to the DRL method shows the SAC method the DDRL employed performs much better than the other two DRL methods. SAC method achieves the highest Dice, Recall, and KAP, but the lowest 95HD compared with the other two commonly used DRL methods.

achieved by SE-based methods is at least 10% higher than that achieved by the DAN. The uncertainty-evaluation further improves the segmentation accuracy further. For instance, regarding the Recall value, the SE methods which combined with uncertainty-evaluation (MCSE and USSE) is about 4% higher than the TCSE, thus uncertainty-estimation is necessary for such weakly-supervised segmentation tasks. For the two SE methods integrated with uncertainty-estimation, our USSE obtains higher segmentation precision than the MCSE, where the Dice is improved by about 3%. This demonstrates the Multi-scale Uncertainty-estimation (MU) which introduces multi-scale attentions into uncertainty-estimation is more capable to evaluate the segmentation reliability and improve the segmentation accuracy than the Monte-Carlo (MC) Uncertainty-estimation.

Evaluation of inter-comparison

Based on Tab. 3.2, Fig. 3.10, and Fig. 3.13, WSTS has the ability to exploit box-level-labeled data and performs much better than fully-supervised methods under the same number pixel-level-labeled data, as well as approaches the fully-supervised methods that are trained with twice number of pixel-level-labeled data. Under 200 patient pixel-level-labeled training data,

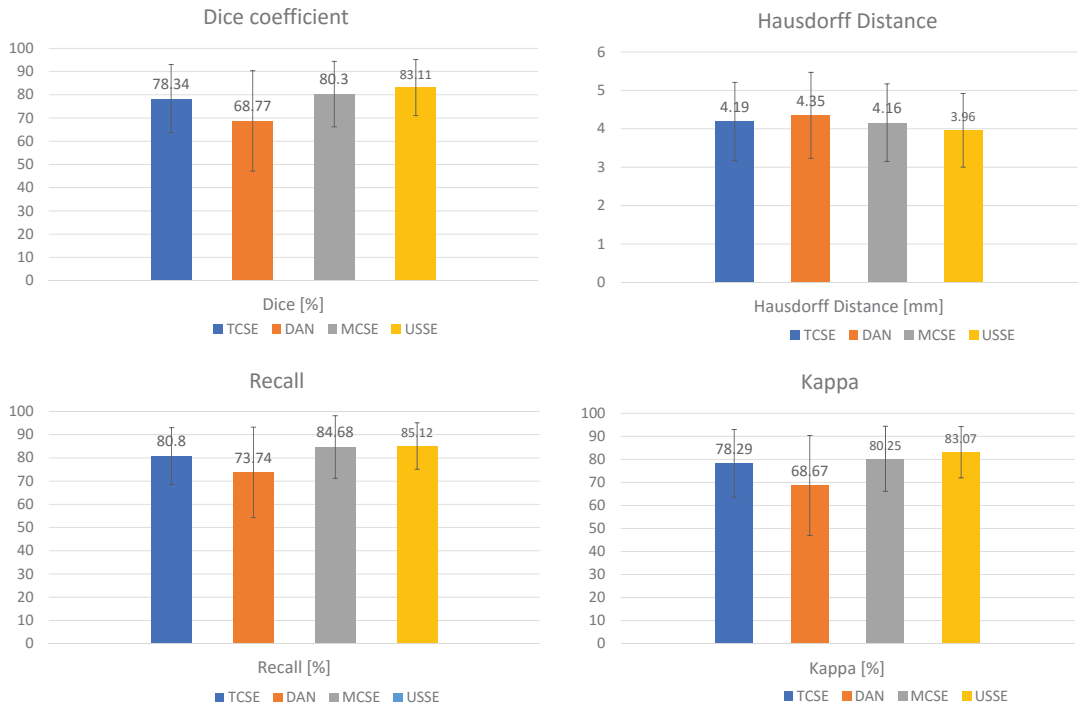


Figure 3.12. Ablation experimental result related to semi-supervised segmentation methods shows the USSE achieves the best performance. Our USSE that deploys Multi-scale Uncertainty-estimation achieves the highest Dice, Recall and KAP, but the lowest 95HD compared with other two state-of-art methods. In addition, SE-based methods (TCSE, MCSE, and USSE) compared with DAN performs better in the tumor segmentation task. The uncertainty-evaluation (MCSE and USSE) further improves the segmentation accuracy further. The Multi-scale Uncertainty-estimation method (USSE) is more capable to improve the segmentation accuracy than the Monte-Carlo Uncertainty-estimation method (MCSE).

U-Net achieves 73.31 ±12.19% in Dice, 4.16 ± 1.10 mm in 95HD, 83.44 ± 10.81% in Recall, and 73.30 ± 12.21% in KAP without the assisting of the enhanced image during training. Under the same condition, Mask R-CNN achieves 87.43±10.42% in Dice, 3.82±0.93 mm in 95HD, 89.55±11.31% in Recall, and 87.41 ± 10.43% in KAP. With the enhanced image as guidance during the training, RgGAN achieves 88.19 ± 10.57% in Dice, 3.72 ± 0.89 mm in 95HD, 89.76 ± 12.04% in Recall and 88.16 ± 10.58% in KAP. With the same condition as RgGAN, PSCGAN achieves 89.31 ± 9.30% in Dice, 3.74 ± 0.87 mm in 95HD, 91.02 ± 9.83% in Recall, and 89.30 ± 9.33% in KAP. Compared with two methods without enhanced image assisting during training, RgGAN and PSCGAN perform much better, which also proves the necessity that exploiting enhanced images to guide the tumor segmentation from non-enhanced images. With half number of pixel-level-labeled data (100 patients) and half number of box-level-labeled data (100 patients), WSTS achieves 83.11±12.11% in Dice, 3.96 ± 0.96 mm in 95HD, 85.12 ± 12.46% in Recall, and 83.07 ± 11.12 in KAP. WSTS's performance approaches

Table 3.2. Inter-comparison experimental result.

Method	Enhanced assisting in training	label used		Metrics			
		pixel -level	box -level	Dice[%]	95HD[mm]	Recall[%]	KAP[%]
U-Net	No	100	0	66.85±22.87	4.35±1.16	73.83±23.31	66.73±22.88
Mask R-CNN	No	100	0	65.70±15.36	4.42±1.02	70.60±15.17	65.65±15.37
RgGAN	Yes	100	0	69.08±18.67	4.32±1.06	75.71±20.40	69.07±18.69
PSCGAN	Yes	100	0	68.11±16.28	4.37±1.03	71.38±17.95	68.04±16.30
U-Net	No	200	0	73.31±12.19	4.16±1.01	83.44±10.81	73.30±12.21
Mask R-CNN	No	200	0	87.43±10.42	3.82±0.93	89.55±11.31	87.41±10.43
RgGAN	Yes	200	0	88.19±10.57	3.72±0.89	89.76±12.04	88.16±10.58
PSCGAN	Yes	200	0	89.31±9.30	3.74±0.87	91.02±9.83	89.30±9.33
WSTS	Yes	100	100	**83.11±12.11**	**3.96±0.96**	**85.12±12.46**	**83.07±11.12**

these fully-supervised methods and even better than U-Net. Fig. 3.13 illustrates some tumor segmentation results from the above methods. Where we can get following findings: 1) when the tumor presents high contrast, five methods segment the tumor successfully; 2) when the tumor is low-contrast and background tissue is complex, RgGAN, and U-Net are tend to be misled by background; 3) the prediction of our method is usually small than the ground-truth, while the predictions of PSCGAN are larger than the ground-truth, which is the reason that PSCGAN achieves a high Recall value.

Under half pixel-level-labeled training data (100 patients), U-Net achieves 66.85 ±22.87% in the Dice, 4.35 ± 1.16 mm in the 95HD, 73.83 ± 23.3‘% in the Recall, and 66.73 ± 22.88% in the KAP without the assisting of the enhanced image during training. Under the same condition, Mask R-CNN achieves 65.70 ± 15.36% in the Dice, 4.42 ± 1.02 mm in the 95HD, 70.60 ± 15.17% in the Recall, and 65.65 ± 15.37%. Under the assistance of the enhanced image during the training stage, RgGAN achieves 69.08 ± 18.67% in the Dice, 4.32 ± 1.06 mm in the 95HD, 75.71 ± 20.40% in the Recall, and 69.07 ± 18.69% in the KAP. PSCGAN achieves 68.11 ± 16.28% in the Dice, 4.37 ± 1.03 mm in the 95HD, 71.38 ± 17.95% in the Recall, and 68.04 ± 16.30% in the KAP. The above results prove that exploiting the enhanced image is important to improve the segmentation accuracy in such tasks, and exploiting the additional box-level-labeled data promotes the segmentation performance greatly. Fig. 3.10 right part also proves the superiority of our method, where the AUC of our method is much larger than the AUCs of the other three methods.

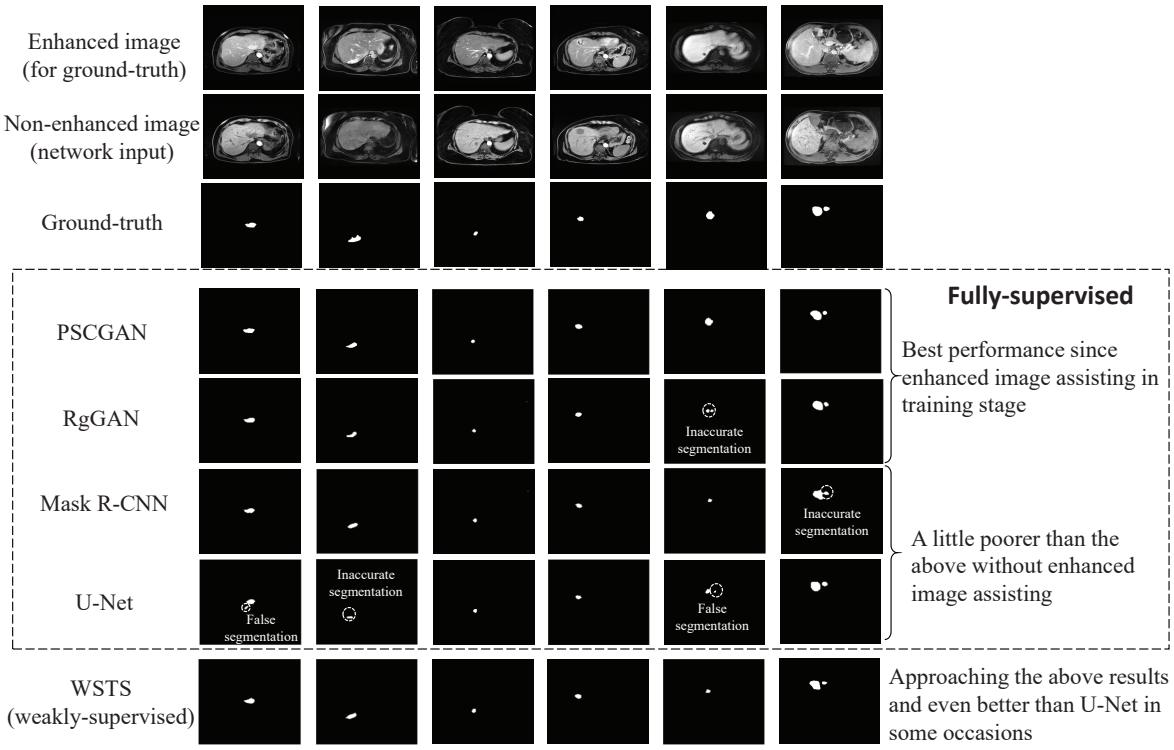


Figure 3.13. The tumor segmentation comparison of the WSTS and other methods shows the performance of WSTS approaches these fully-supervised methods and even better than the U-Net. The first and second row illustrates enhanced images and non-enhanced images respectively, the third row shows the ground-truth, and the rest rows illustrate the segmentation of five methods. By exploiting box-level-labeled data, WSTS achieves the segmentation result that approaching the segmentation of fully-supervised methods (PSCGAN, RgGAN and Mask R-CNN), and even better than U-Net.

References

[1] A. Radtke, S. Nadalin, G. Sotiropoulos, E. Molmenti, T. Schroeder, C. Valentin-Gamazo, H. Lang, M. Bockhorn, H. Peitgen, C. Broelsch, *et al.*, "Computer-assisted operative planning in adult living donor liver transplantation: a new way to resolve the dilemma of the middle hepatic vein," *World journal of surgery*, vol. 31, no. 1, p. 175, 2007.

[2] J. Chapiro, R. Duran, M. Lin, R. E. Schernthaner, Z. Wang, B. Gorodetski, and J.-F. Geschwind, "Identifying staging markers for hepatocellular carcinoma before transarterial chemoembolization: comparison of three-dimensional quantitative versus non–three-dimensional imaging markers," *Radiology*, vol. 275, no. 2, pp. 438–447, 2015.

[3] B. Nordlinger, M. Guiguet, J.-C. Vaillant, P. Balladur, K. Boudjema, P. Bachellier, and D. Jaeck, "Surgical resection of colorectal carcinoma metastases to the liver: a prognostic scoring system to improve case selection, based on 1568 patients," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 77, no. 7, pp. 1254–1262, 1996.

[4] M. Moghbel, S. Mashohor, R. Mahmud, and M. I. B. Saripan, "Automatic liver segmentation on computed tomography using random walkers for treatment planning," *EXCLI journal*, vol. 15, p. 500, 2016.

[5] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[6] E. A. Sadowski, L. K. Bennett, M. R. Chan, A. L. Wentland, A. L. Garrett, R. W. Garrett, and A. Djamali, "Nephrogenic systemic fibrosis: risk factors and incidence estimation," *Radiology*, vol. 243, no. 1, pp. 148–157, 2007.

[7] C. Xu, L. Xu, Z. Gao, S. Zhao, H. Zhang, Y. Zhang, X. Du, S. Zhao, D. Ghista, H. Liu, *et al.*, "Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture," *Medical image analysis*, vol. 50, pp. 82–94, 2018.

[8] F. Stacul, A. J. van der Molen, P. Reimer, J. A. Webb, H. S. Thomsen, S. K. Morcos, T. Almén, P. Aspelin, M.-F. Bellin, O. Clement, *et al.*, "Contrast induced nephropathy: updated esur contrast media safety committee guidelines," *European radiology*, vol. 21, no. 12, pp. 2527–2541, 2011.

[9] M. A. Ibrahim, B. Hazhirkarzar, and A. B. Dublin, "Magnetic resonance imaging (mri) gadolinium," in *StatPearls [Internet]*, StatPearls Publishing, 2020.

[10] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *International Conference on Information Processing in Medical Imaging*, pp. 554–565, Springer, 2019.

[11] X. Xiao, J. Zhao, Y. Qiang, J. Chong, X. Yang, N. G.-F. Kazihise, B. Chen, and S. Li, "Radiomics-guided gan for segmentation of liver tumor without contrast agents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 237–245, Springer, 2019.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[13] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu, "An artificial agent for anatomical landmark detection in medical images," in *International conference on medical image computing and computer-assisted intervention*, pp. 229–237, Springer, 2016.

[14] G. Maicas, G. Carneiro, A. P. Bradley, J. C. Nascimento, and I. Reid, "Deep reinforcement learning for active breast lesion detection from dce-mri," in *International conference on medical image computing and computer-assisted intervention*, pp. 665–673, Springer, 2017.

[15] X. Han, "Automatic liver lesion segmentation using a deep convolutional neural network method," *arXiv preprint arXiv:1704.07239*, 2017.

[16] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[17] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," *arXiv preprint arXiv:1808.03887*, 2018.

[18] D. Kainmüller, T. Lange, and H. Lamecker, "Shape constrained automatic segmentation of the liver based on a heuristic intensity model," in *Proc. MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge*, pp. 109–116, 2007.

[19] R. Beichel12, C. Bauer, A. Bornik, E. Sorantin, and H. Bischof, "Liver segmentation in ct data: A segmentation refinement approach," *Proceedings of" 3D Segmentation in The Clinic: A Grand Challenge*, pp. 235–245, 2007.

[20] A. Bornik, R. Beichel, E. Kruijff, B. Reitinger, and D. Schmalstieg, "A hybrid user interface for manipulation of volumetric medical data," in *3D User Interfaces (3DUI'06)*, pp. 29–36, IEEE, 2006.

[21] A. Beck and V. Aurich, "Hepatux-a semiautomatic liver segmentation system," *3D Segmentation in The Clinic: A Grand Challenge*, pp. 225–233, 2007.

[22] B. M. Dawant, R. Li, B. Lennon, and S. Li, "Semi-automatic segmentation of the liver and its evaluation on the miccai 2007 grand challenge data set," *3D Segmentation in The Clinic: A Grand Challenge*, pp. 215–221, 2007.

[23] J. Lee, N. Kim, H. Lee, J. B. Seo, H. J. Won, Y. M. Shin, and Y. G. Shin, "Efficient liver segmentation exploiting level-set speed images with 2.5 d shape propagation," in *Proceedings of the MICCAI Workshop on 3-D Segmentat. Clinic: A Grand Challenge*, pp. 189–196, 2007.

[24] A. Wimmer, G. Soza, and J. Hornegger, "Two-stage semi-automatic organ segmentation framework using radial basis functions and level sets," *3D segmentation in the clinic: a grand challenge*, pp. 179–188, 2007.

[25] P. Slagmolen, A. Elen, D. Seghers, D. Loeckx, F. Maes, and K. Haustermans, "Atlas based liver segmentation using nonrigid registration with a b-spline transformation model," in *Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge*, pp. 197–206, 2007.

[26] W. Huang, Y. Yang, Z. Lin, G.-B. Huang, J. Zhou, Y. Duan, and W. Xiong, "Random feature subspace ensemble based extreme learning machine for liver tumor detection and segmentation," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4675–4678, IEEE, 2014.

[27] X. Zhang, J. Tian, D. Xiang, X. Li, and K. Deng, "Interactive liver tumor segmentation from ct scans using support vector classification with watershed," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6005–6008, IEEE, 2011.

[28] C.-L. Kuo, S.-C. Cheng, C.-L. Lin, K.-F. Hsiao, and S.-H. Lee, "Texture-based treatment prediction by automatic liver tumor segmentation on computed tomography," in

2017 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 128–132, IEEE, 2017.

[29] G. Sethi, B. S. Saini, and D. Singh, "Segmentation of cancerous regions in liver using an edge-based and phase congruent region enhancement method," *Computers & Electrical Engineering*, vol. 53, pp. 244–262, 2016.

[30] S. Patil, V. Udupi, and D. Patole, "A robust system for segmentation of primary liver tumor in ct images," *International Journal of Computer Applications*, vol. 75, no. 13, 2013.

[31] A. Hoogi, C. F. Beaulieu, G. M. Cunha, E. Heba, C. B. Sirlin, S. Napel, and D. L. Rubin, "Adaptive local window for level set segmentation of ct and mri liver lesions," *Medical image analysis*, vol. 37, pp. 46–55, 2017.

[32] T. AMARAJOTHI, S. MANIKANDAN, and K. MUTHUKKUTTI, "Liver tumor segmentation using single level set method with shape and intensity prior," *International Journal of Applied Engineering Research*, vol. 10, no. 20, 2015.

[33] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, *et al.*, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 415–423, Springer, 2016.

[34] C. Sun, S. Guo, H. Zhang, J. Li, M. Chen, S. Ma, L. Jin, X. Liu, X. Li, and X. Qian, "Automatic segmentation of liver tumors from multiphase contrast-enhanced ct images based on fcns," *Artificial intelligence in medicine*, vol. 83, pp. 58–66, 2017.

[35] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[36] A. Alansary, L. Le Folgoc, G. Vaillant, O. Oktay, Y. Li, W. Bai, J. Passerat-Palmbach, R. Guerrero, K. Kamnitsas, B. Hou, *et al.*, "Automatic view planning with multi-scale deep reinforcement learning agents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 277–285, Springer, 2018.

[37] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," 2011.

[38] T. Dai, M. Dubois, K. Arulkumaran, J. Campbell, C. Bass, B. Billot, F. Uslu, V. de Paola, C. Clopath, and A. A. Bharath, "Deep reinforcement learning for subpixel neural tracking," in *International Conference on Medical Imaging with Deep Learning*, pp. 130–150, 2019.

[39] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.

[40] S. Guan, A. Khan, S. Sikdar, and P. Chitnis, "Fully dense unet for 2d sparse photoacoustic tomography artifact removal," *IEEE journal of biomedical and health informatics*, 2019.

[41] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613, Springer, 2019.

[42] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9080–9089, 2018.

[43] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, pp. 1195–1204, 2017.

[44] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[45] 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.

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

[47] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[48] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 408–416, Springer, 2017.

[49] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[50] C. Xu, L. Xu, P. Ohorodnyk, M. Roth, B. Chen, and S. Li, "Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal gans," *Medical Image Analysis*, p. 101668, 2020.

CHAPTER 4

The last chapter of this thesis reviews the motivation and research objectives and summarizes the important findings and conclusions of Chapter 2 and Chapter 3, as well as discuss their impacts on the field. The limitation of the work presented in this thesis is discussed along with its inspiration regarding deep reinforcement learning to future studies.

4 CONCLUSION AND FUTURE DIRECTIONS

4.1 Overview of Rationale and Research Questions

Artificial Intelligence (AI), as one of the most innovative technologies in the new century, has changed the human lifestyle significantly and brought great convenience to medical image analysis. Deep Learning (DL), as the foundation of AI technology, is applied to automatic medical image analysis is to provide clinicians a computer-assisting tool to diagnose diseases. Medical object detection and segmentation, as the principal and initial step preceding other image analysis tasks, is the targeted application of DL methods. Medical object detection proposes the region of the desirable object in the medical image, thus it can provide the object's location. Medical image segmentation classifies every pixel into the desirable object or the background, thus it can provide several properties of the desirable object, such as the object's size and boundary.

Although DL has evolved substantially, it also exposes several shortcomings. DL methods completely rely on engineers and blindly executes the solution, they are decoupled from the understanding of the task at hand. Because DL methods have no ability to discover the intrinsic knowledge about the task, they often suffer from computational limitations, sub-optimal parameter optimization, and weak generalization due to over-fitting.

The design of leveraging Deep Reinforcement Learning (DRL) to segment medical images has a great potential to address the limitations of traditional DL methods. By combining the strong decision-making ability of RL and the powerful perception of DL, DRL has a sequential process with changing attention-focusing that gradually accumulate evidence of certainty when searching the desirable object. Therefore, in object detection and segmentation tasks, with the sequential process, DRL can simultaneously both searching strategy and the appearance of the object of interest to determine the object location. With the object location as a prior, the detection and segmentation method can focus on the desirable object and obtain more accurate results. Compared with the traditional DL method which directly predicts the object location

with a mapping according to appearance, DRL can avoid falling into sub-optimal caused by the repetitive or low-contrast object appearance, thereby determining the object location effectively.

However, the medical object detection and segmentation cover a large and complex scope, and no DRL methods have been employed in this field. Therefore, from simple to complex, this thesis deploys DRL into two challenging and representative medical object detection and segmentation tasks. The particular research objectives for the above two tasks include: 1) Modeling the spine anatomy as a sequential decision-making process and Leveraging DRL to facilitate accurate vertebral body segmentation (Chapter 2). 2) Utilizing the policy (strategy) developed by DRL as guidance to determine the tumor location in the non-contrast-enhanced liver image and then achieving liver tumor segmentation (Chapter3).

4.2 Summary and Conclusions

In Chapter 2, we developed a novel Sequential Conditional Reinforcement Learning network (SCRL) to automatically detect and segment vertebral bodies from MRI images. Unlike the existing techniques, SCRL models the sequential correlation of VBs with DRL and focuses segmentation and detection on each desirable VB to obtain the final result. SCRL for the first time leverages powerful DRL to exploit the anatomy of the spine to process VBs, thereby handling the complex background in the spine image, as well as pathological variations or anatomic variants of VBs. The experimental results demonstrate the effectiveness of SCRL to achieve remarkable detection and segmentation accuracy (with overall 92.3% IoU in VB detection, 92.6% Dice in VB segmentation and 96.4% mean accuracy in VB classification). These results demonstrate that SCRL can be an efficient and accurate computer aided-diagnostic tool to assist clinicians when diagnosing spine diseases.

In Chapter 3, we proposed a Weakly-Supervised Teacher-Student network (WSTS) to segment the liver tumor from non-enhanced images. WSTS creatively deploys a weakly-supervised teacher-student framework to exploit the visible tumor feature in the enhanced image as guidance and assist the segmentation in the non-enhanced images. In particular, WSTS proposes DDRL to exploit the tumor spatial feature in the enhanced image and transfer the feature via DRL strategies to guide the tumor detection in the non-enhanced image. WSTS also proposes the USSE, which predicts tumor masks for the box-level-labeled enhanced image, thus the mask is transferred as the tumor shape feature to improve tumor segmentation in the non-enhanced image. The experiment confirms the effectiveness of WSTS, where WSTS achieves

83.11% in Dice and 85.12% in Recall with the data from 200 patients (100 patients with pixel-level label and 100 patients with box-level label). These results demonstrate WSTS can be an efficient computer aided-diagnostic tool to assist clinicians when diagnosing liver tumors from non-enhanced images.

4.3 Significance and Impact

As the development of DL-based methods, the limitations of DL in the medical object detection and segmentation gradually reveal. The traditional mapping-establishment based on CNN cannot meet the clinical expectation in medical object detection and segmentation. Thus, researchers are seeking a breakthrough to obtain more excellent performance in this field. DRL as the newest AI technology has a totally-different principle as DL, it has shown its competence in many other medical image analysis fields, such as in medical object detection. In medical image analysis, DRL has a cognitive-like process, so that it has great potential to inject new ideas to medical object detection and segmentation and address the limitations of traditional DL methods.

This thesis has validated that deploying DRL in the VB detection and segmentation, as well as the liver tumor segmentation achieved better performance than traditional DL-based methods. For the VB segmentation, the proposed method for the first time modeled the spine anatomy as a sequential decision-processing by taking advantage of the spine spatial architecture. It offers researchers a new idea to address medical image analysis through modeling the anatomy with DRL. For the non-enhanced liver tumor segmentation, the proposed method for the first time to transfer tumor knowledge from the enhanced image to the non-enhanced image through DRL strategy. It provides researchers a new solution to address the tumor segmentation in the non-enhanced image by transferring knowledge from the enhanced image.

4.4 Limitations

4.4.1 Study Specific Limitations

Chapter 2: Sequential Conditional Reinforcement Learning for Simultaneous Vertebral Body Detection and Segmentation with Modeling the Spine Anatomy

In Chapter 2, the proposed SCRL algorithm analyzed VBs along the spine and ended when the classification branch outputs the terminal signal. Such a mechanism makes the VB classification accuracy depends heavily on the accuracy classification branch. Particularly, if the

classification branch outputs the terminal signal too early (i.e., mis-classifies a lumbar VB as the sacrum) or too late (i.e, mis-classifies the sacrum as a lumbar VB), all the VBs in this image will be mis-classified. Thus, It is a limitation that overall classification accuracy relies heavily on the accuracy of the classification branch.

Chapter 3: Weakly-Supervised Teacher-Student Network for Liver Tumor Segmentation from Non-enhanced Images

In Chapter 3, the proposed WSTS algorithm conveyed tumor spatial features from the enhanced image to the non-enhanced image through the strategy, which is effective but still has improvable space. In WSTS, the tumor-approaching strategy was regarded as a high-level understanding of DRL to liver image, thus it was taken as guidance to teach the tumor-detection in the non-enhanced image, which is a kind of knowledge-transfer. Although the experiment validated such a knowledge-transfer is effective, it is a sub-optimal way. Because as Deep Transfer Reinforcement Learning, some advanced algorithms have been proposed. Integrating these new algorithms into knowledge-transfer may achieve better performance.

4.4.2 General Limitation

The research that applying DRL in medical object detection and segmentation possesses a limitation is that the object-area-proposal of DRL and the object-analysis (detection and segmentation) of DL were optimized separately, which requires more experiments to tune their hyper-parameters individually. In detail, in the current setting, the object-area-proposal of DRL and the object-analysis of DL were firstly trained-separately, and then they were integrated together to fine-tune the object-analysis of DL. Therefore, it took a longer training period and more training resources to obtain experiment results.

4.5 Future Directions

4.5.1 End-to-end framework design

For simplifying the training procedure and achieve better performance, it is highly-desirable to design an end-to-end framework and integrate DRL-based object-proposal and DL-based object analysis (detection and segmentation) together. In detail, in the end-to-end framework, the reward should be allocated by the performance of detection and segmentation instead of being allocated blindly according to the ground-truth. Such an end-to-end framework only needs to be trained and tuned once instead of being training separately and then fine-tuning

together. This would reduce the training time and training resources greatly. More importantly, an end-to-end framework helps the machine to understand the whole medical object detection and segmentation task instead of the only object-area proposal, which may facilitate the model to achieve better performance compared with current setting.

4.5.2 Three-dimensional (3D) image segmentation

To meet future clinical needs, we seek to extend our DRL-based segmentation method to 3D image applications. Since our object-region proposal in DRL performs action prediction with four variables (center [x, y] and size [w, h]), the 3D object detection and segmentation would increase the number of prediction variables (center [x, y, z] and size [w, h, l]). Such a change would increase the demand for the training data and the training resource. However, if successful, it would provide clinicians with more comprehensive information about the object of interest.

4.5.3 Extension to other tasks

Considering this thesis has validated the effectiveness of DRL in two representative medical object detection and segmentation tasks, it would be achievable to extend DRL to other medical object detection and segmentation tasks, such as organ detection and segmentation. In detail, the VB segmentation in this thesis modeled the spine anatomy, which indicates that other tasks with special anatomical architecture can also deploy our DRL method to achieve great segmentation performance. For instance, organs have a relatively-fixed location and size, it is easy for DRL to model their spatial knowledge and propose their area accurately. The liver tumor segmentation of non-enhanced images in this thesis leverages DRL for knowledge-transferring, which can be extended to other segmentation tasks with uncertain location and size. For instance, cancers usually have various locations and sizes because of pathology, it is still achievable to adapt our DRL method in such a task.

APPENDIX

APPENDIX A: Derivation of Back Propagation

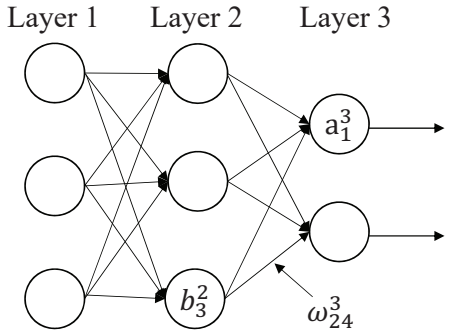


Figure A.1. An ANN example for back propagation.

Fig. A.1 is a three-layer ANN, where w_{jk}^l represents the weight of the k-th neuron in $l-1$ layer connected to the j-th neuron in l layer. b_j^l represents the bias of the j-th neuron in l layer. z_j^l represents the input of the j-th neuron in l layer, namely:

$$z_j^l = \sum_k \omega_{jk}^l a_k^{l-1} + b_j^l \tag{A.1}$$

a_j^l is the output of the j-th neuron in l layer, namely:

$$a_j^l = \sigma(\sum_k \omega_{jk}^l a_k^{l-1} + b_j^l) \tag{A.2}$$

where σ is an activation function. We assume C is the error between the network output and the true label. Thus the error caused by the j-th neuron in l layer is defined as:

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} \tag{A.3}$$

The error caused by the last layer is:

$$\delta^L = \nabla_a C \bigodot \sigma'(z^L) \tag{A.4}$$

where $\bigodot$ is Hadamard product for the point-to-point product between matrices (or vectors). The derivation of Eq. A.4 is:

$$\because \delta_j^L = \frac{\partial C}{\partial z_j^L} = \frac{\partial C}{\partial a_j^L} \cdot \frac{\partial a_j^L}{\partial z_j^L} \tag{A.5}$$

$$\therefore \delta^L = \frac{\partial C}{\partial a_j^L} \bigodot \frac{\partial a_j^L}{\partial z_j^L} = \nabla_a C \bigodot \sigma'(z^L) \tag{A.6}$$

From back to front, the error caused by every layer is:

$$\delta^l = ((\omega^{l+1})^T \delta^{l+1}) \bigodot \sigma'(z^l) \tag{A.7}$$

The derivation of Eq. A.7 is:

$$\because \delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j^l} \cdot \frac{\partial a_j^l}{\partial z_j^l} \tag{A.8}$$

$$= \sum_k \delta_k^{l+1} \cdot \frac{\partial (w_{kj}^{l+1} a_j^l + b_k^{l+1})}{\partial a_j^l} \cdot \sigma'(z_j^l) \tag{A.9}$$

$$= \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \cdot \sigma'(z_j^l) \tag{A.10}$$

$$\therefore \delta^l = ((\omega^{l+1})^T \delta^{l+1}) \bigodot \sigma'(z^L) \tag{A.11}$$

The gradient of the error C on the weight ω_{jk}^l is:

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \tag{A.12}$$

The derivation of Eq. A.12 is:

$$\frac{\partial C}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial w_{jk}^l} = \delta_j^l \cdot \frac{\partial \left(w_{jk}^l a_k^{l-1} + b_j^l \right)}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \tag{A.13}$$

The gradient of the error C on the bias b_j^l is:

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \tag{A.14}$$

The derivation of Eq. A.14 is:

$$\frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial b_j^l} = \delta_j^l \cdot \frac{\partial \left(w_{jk}^l a_k^{l-1} + b_j^l \right)}{\partial b_j^l} = \delta_j^l \tag{A.15}$$

Thus, according to the gradient descent, the weight and bias can be trained:

$$\begin{aligned} w^l &\rightarrow w^l - \frac{\eta}{m} \sum_x \delta^{x,l} \left(a^{x,l-1}\right)^T \\ b^l &\rightarrow b^l - \frac{\eta}{m} \sum_x \delta^{x,l} \end{aligned} \tag{A.16}$$

APPENDIX B: Training process of Deep Q-learning

[1] Initialize replay memory D
[2] Initialize state-action value function Q with random weights θ
[3] Initialize target state-action value function $\hat{Q}$ with weight θ^{-}
[4] For episode = 1, M **do**
[5] Initialize sequence s_1
[6] For t = 1, T **do**
[7] With probability η select a random action a_t
[8] Otherwise select $a_t = argmax_a Q(s_t, a|\theta)$
[9] Execute action a_t in emulartor and receive reward r_t and observe new state s_{t+1}
[10] Store transition (s_t, a_t, r_t, s_{t+1}) in D, set $s_t = s_t$
[11] Sample random minibatch of transitions (s_j, a_j, r_j, s_{j+1}) from D
[12] Set $y_i = \begin{cases} r_j, & \text{if episode terminates at step } j+1 \\ r_j + \gamma \max_{a'} \hat{Q}\left(\phi_{j+1}, a'|\theta^{-}\right), & \text{otherwise} \end{cases}$
[13] Perform a gradient descent step on $\left(y_j - Q\left(\phi_j, a_j; \theta\right)\right)^2$ with respect to the network parameters θ
[14] Every C steps reset $\hat{Q} = Q$
[15] **End For**
[16] **End For**

APPENDIX C: Proof of $P(s|i) \propto i_s^{-1}$

Assuming: in an episode, an agent selects actions to approach an target and increase accuracy: $Q = [S, A, I]$; Where S represents various states, $S = \{s_t, s_{t-1}, ..., s_1\}$; A represents the action corresponding to S, $A = \{a_t, a_{t-1}, ..., a_1\}$; I represents the accuracy corresponding to S, $I = \{i_t, i_{t-1}, ..., i_1\}$. According to the state transition relation, the probability of s_t can be achieved is:

$$\begin{aligned} P(s_t, i_t) &= P(s_{t-1} \otimes a_{t-1}, i_t) \\ &= P(s_{t-1}, i_{t-1}) * P(a_{t-1}|s_{t-1}) \end{aligned} \tag{C.1}$$

Therefore, for the whole episode, the joint probability distribution of Q is:

$$P(s_t, i_t, s_{t-1}, a_{t-1}, ..., s_1, a_1) = P(s_1, i_1) \prod_{j=1}^{t-1} P(a_j|s_j) \tag{C.2}$$

In equation (A.2), $\{P(a_i|s_i), i \in [1, t-1]\}$ is a reflection of the decision-making ability of DRL. Therefore:

$$\begin{cases} P(s_1, i_1) = 1 \\ 0 \leq P(a_i|s_i) < 1 \end{cases} \tag{C.3}$$

With equation (A.2), (A.3) for one episode is obtained:

$$P(s_t, i_t, s_{t-1}, a_{t-1}, ..., s_1, a_1) < P(s_{t-1}, i_{t-1}, s_{t-2}, a_{t-2}, ..., s_1, a_1) < ... < P(s_1, i_1) \tag{C.4}$$

Since the training data is sampled with lots of episodes, the expected probability of a state s that gets reward i_j in N episodes is:

$$P(s|i_j) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{(s_t, i_t) \sim Q_k}^{i_t = i_j} P(s_t, i_t) \tag{C.5}$$

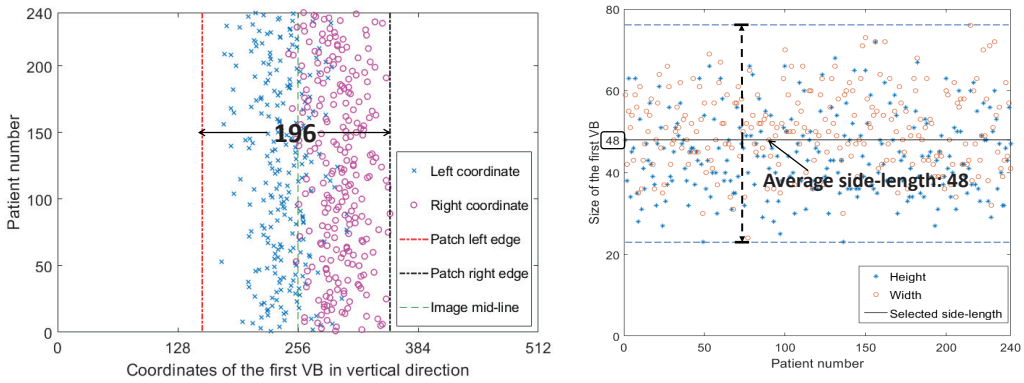
where j indicates the value of reward i, i.e., $i_j > i_{j-1} > ... > i_1$. According to equation (A.3)(A.4)(A.5) and the Law of Large Numbers (LLN), the following conclusion is drawn:

$$\begin{cases} P(s^*|i_j) < P(s^*|i_{j-1}) < ... < P(s^*|i_1) \\ i_j > i_{j-1} > ... > i_1 \end{cases} \tag{C.6}$$

Thus the probability of the existence of a state s is negatively related to the accuracy i that state s achieves:

$$P(s|i) \propto i_s^{-1} \tag{C.7}$$

APPENDIX D: Distributions for AMRL hyperparameter setting



a) The coordinates of the first VB for every patient

b) The size of the first VB for every patient

Figure D.1. Left figure: The vertical direction coordinates of the first VB: the blue "x" and pink circle represent the left and right coordinates of the first VB respectively; the dotted red and black lines represent the coordinate of our first image patch (196*196). Right figure: The side length of the first VB, where blue star is height and the orange circle is width. After counting, the average side-length of the first VB is 48.

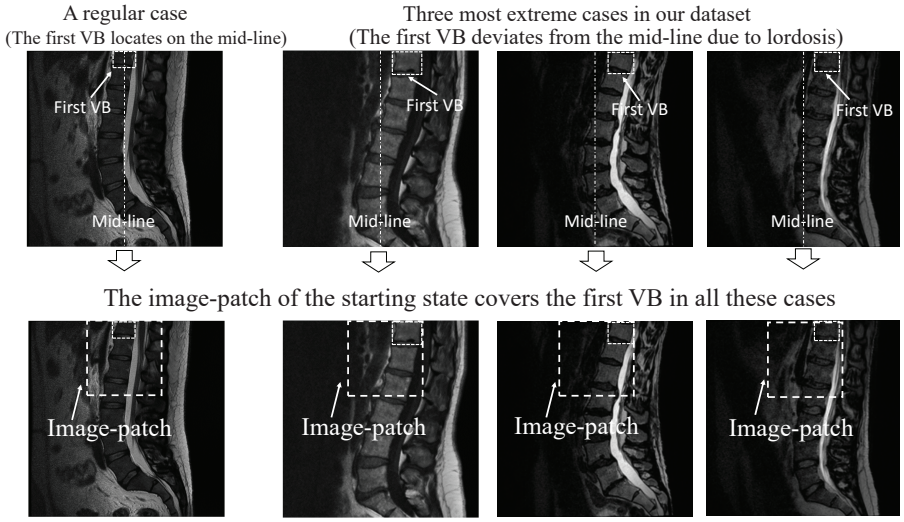


Figure D.2. Some examples to demonstrate that the image-patch in the starting state can cover the first VB totally.

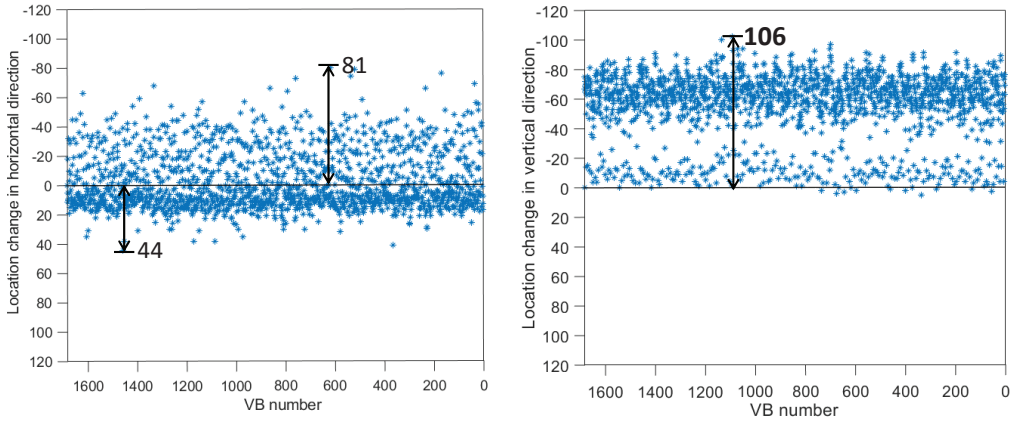


Figure D.3. The location changes between adjacent VBs in vertical and horizontal direction. The maximum location change between adjacent VBs is 106 in vertical direction.

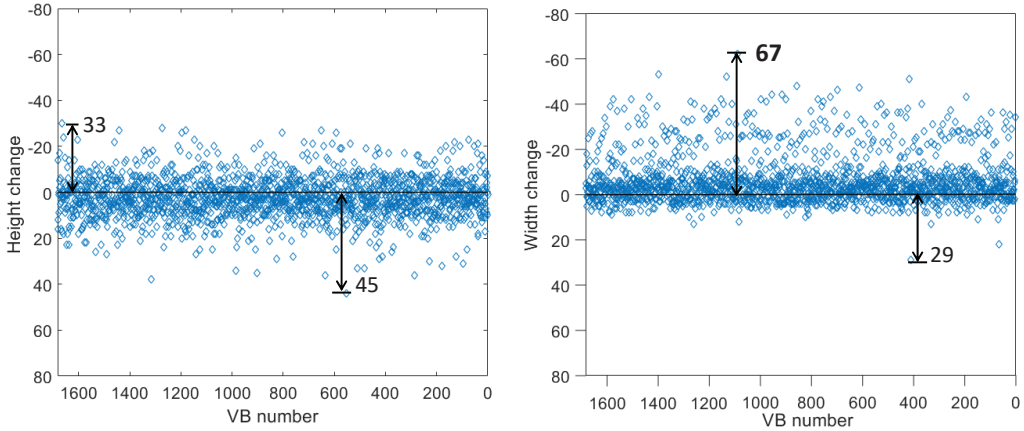


Figure D.4. The size changes between adjacent VBs in vertical and horizontal direction. The maximum size change is 67 in width.

APPENDIX E: Derivation from Eq. 3.10 to Eq. 3.11

We first rewrite the Eq. 3.10 as follows,

$$\pi^* = \underset{\pi}{argmax} \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi} \{r(s_t,a_t) + \alpha[\mathcal{H}(\pi(\cdot|s_t)) + \mathcal{D}_{KL_1}(\pi(\cdot|s_t)||\pi'(\cdot|s_t))]\} \tag{E.1}$$

In this case, the value function $V(s_t)$ is

$$V(s_t) = \mathbb{E}_{a_t\sim\pi}[Q(s_t,a_t) - log\pi(a_t|s_t) + log\frac{\pi(a_t|s_t)}{\pi'(a_t|s_t)}] \tag{E.2}$$

During training V and optimizing the parameter ψ, the gradient of V is

$$J_V(\psi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left\{\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi}[Q_\theta(s_t,a_t) - log\pi(a_t|s_t) + log\frac{\pi(a_t|s_t)}{\pi'(a_t|s_t)}])^2\right\} \tag{E.3}$$

Then, $J_V(\psi)$ can be reformulated:

$$J_V(\psi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left\{\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi_\phi}[Q_\theta(s_t,a_t) - log\pi_\phi(a_t|s_t) + (log\pi_\phi(a_t|s_t) - log\pi'(a_t|s_t))])^2\right\} \tag{E.4}$$

Thus, Eq. 3.11 is obtained:

$$J_V(\psi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left\{\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi_\phi}[Q_\theta(s_t,a_t) - log\pi'(a_t|s_t)]^2\right\} \tag{E.5}$$

APPENDIX F: The networks in DDRL

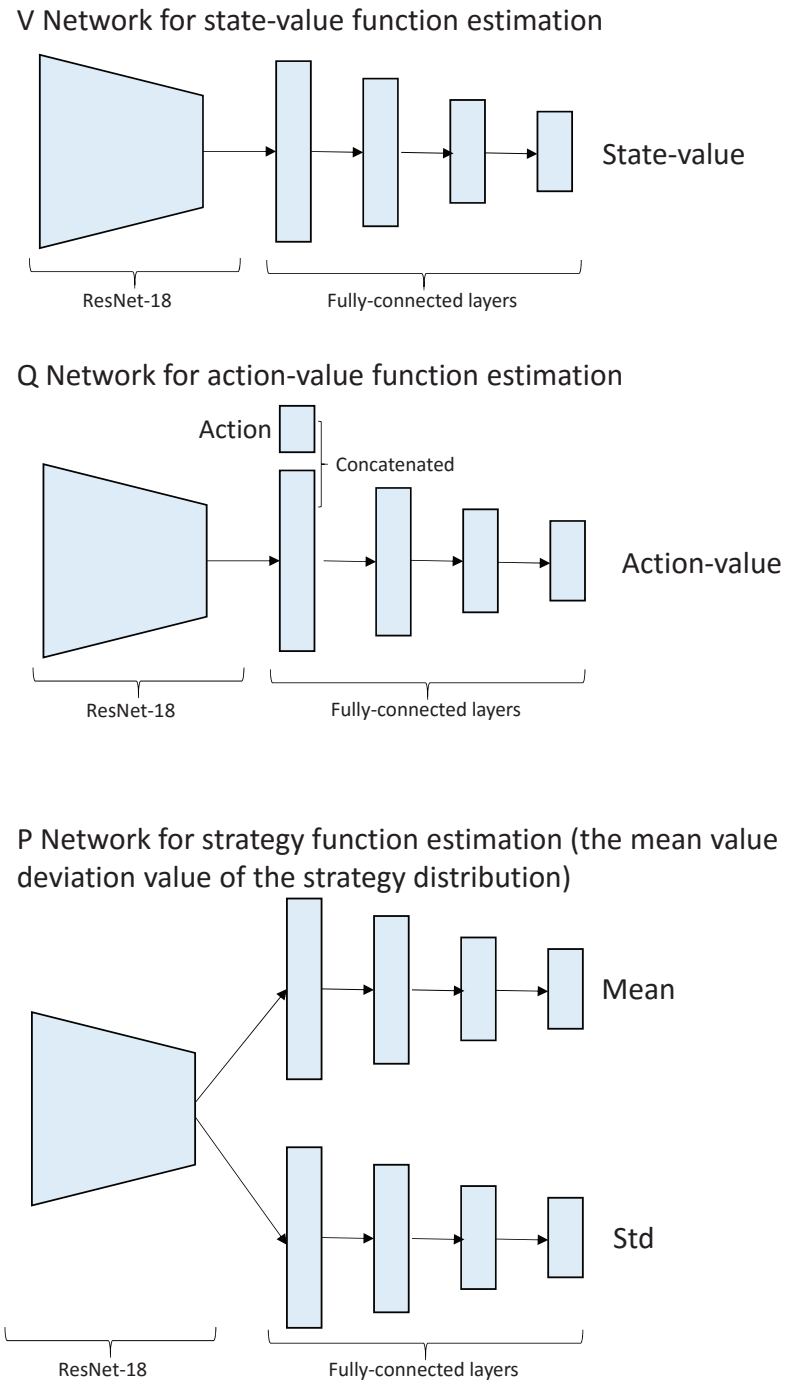


Figure F.1. The network architectures to approximate the state-value function, action-value function, and strategy function

Curriculum Vitae

Name:	Dong Zhang
Post-Secondary Education and Degrees:	Northwestern Polytechnical University Xi'an, Shannxi, China Information and Control 2014-2018 Bachelor of Electrical Engineering University of Western Ontario London, ON Biomedical Engineering, Imaging 2018 - 2020 Masters in Engineering Science
Honours and Awards:	Western Graduate Research Scholarship 2018-2020 Institutional $11, 000/year CAD
Related Work Experience:	Graduate Student Research Assistant The University of Western Ontario London, ON, Canada 2018 - 2020

Publications:

A. Referred Journal Manuscripts (2 Published, 2 Under revisions)

Published (2)

1. Tam CM, **Zhang D**, Chen B, Peters T, Li S. Holistic Multitask Regression Network for Multiapplication Shape Regression Segmentation. Medical Image Analysis. 2020 Jul 11:101783.

2. **Zhang D**, Chen B, Li S, Sequential Conditional Reinforcement Learning for Simultaneous Vertebral Body Detection and Segmentation with Modeling the Spine Anatomy. Medical Image Analysis. 67 (2020): 101861.

Under revisions(2)

1. **Zhang D**, Li S. Weakly-Supervised Teacher-Student Network for Liver Tumor Segmentation from Non-enhanced Images. Medical Image Analysis.

2. Xu C, **Zhang D (Co-First)**, and Li S. Synthesis of Gadolinium-enhanced Liver Tumors on non-enhanced liver MRI Using Interacted Pixel-level Agents Reinforcement learning. Medical Image Analysis.

B. Published Referred Conference Paper (1)

1. Luo G, Dong S, Wang K, **Zhang D**, Gao Y, Chen X, Zhang H, Li S. A Deep Reinforcement Learning Framework for Frame-by-Frame Plaque Tracking on Intravascular Optical Coherence Tomography Image. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 12-20). Springer, Cham.