3-9-2020 10:00 AM

Machine Learning towards General Medical Image Segmentation

Clara Tam
The University of Western Ontario

Supervisor
Li, Shuo
The University of Western Ontario Co-Supervisor
Peters, Terry
The University of Western Ontario

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

Abstract

The quality of patient care associated with diagnostic radiology is proportionate to a physician's workload. Segmentation is a fundamental limiting precursor to diagnostic and therapeutic procedures. Advances in machine learning aims to increase diagnostic efficiency to replace single applications with generalized algorithms. We approached segmentation as a multitask shape regression problem, simultaneously predicting coordinates on an object's contour while jointly capturing global shape information. Shape regression models inherent point correlations to recover ambiguous boundaries not supported by clear edges and region homogeneity. Its capabilities was investigated using multi-output support vector regression (MSVR) on head and neck (HaN) CT images. Subsequently, we incorporated multiplane and multimodality spinal images and presented the first deep learning multiapplication framework for shape regression, the holistic multitask regression network (HMR-Net). MSVR and HMR-Net's performance were comparable or superior to state-of-the-art algorithms. Multiapplication frameworks bridges any technical knowledge gaps and increases workflow efficiency.

Keywords: segmentation, machine learning, deep learning, shape regression, multiapplication, multitask learning

Lay Summary

The development of new image analysis techniques has allowed doctors to better understand the content of an image. Segmentation, a technique to isolate regions of interest, is used in medical interventions such as disease detection, tracking disease progression, and evaluating for surgical procedures, and radiation therapy. The integration of artificial intelligence (AI) technology into medical image analysis can enhance and streamline image segmentation practices. However, existing methods are organ-specific and cannot be adapted. Multitask learning techniques were explored to create a machine learning framework for multiple applications. We approached the segmentation task as a multitask prediction problem (estimating multiple tasks such as organ class, location, and boundary points simultaneously) and introduced shape or boundary point regression; an innovative new technique. Shape regression directly predicts coordinates of an object's shape contour and simultaneously captures its shape. Compared to conventional image pixel segmentation, shape regression can model the natural correlation between points to recover unclear boundaries not supported by clear edges and uniform pixel regions. To determine if a generalized algorithm using shape regression was feasible, we first implemented the technique using a traditional machine learning method, multi-output support vector regression (MSVR). MSVR was applied to head and neck (HaN) CT images consisting of 18 target organs. In another study, we used a more modern machine learning technique, deep learning, for shape regression and expanded the application scope for MR and CT spinal images in both axial and sagittal planes. We presented the first deep learning multiapplication framework for shape regression, the holistic multitask regression network (HMR-Net). MSVR and HMR-Net's performance were similar or superior to current algorithms in literature. The successful performance of an automated multiapplication framework provides many benefits to clinical routine. It bridges any technical knowledge gaps and increases workflow efficiency.

Co-Authorship Statement

The following thesis contains 2 manuscripts: one has been published in a conference proceeding and the other has been submitted to a peer-review journal. Clara Tam, as the first author, was a significant contributor to both studies, data processing, training and validation experiments, data analysis, and manuscript preparation. Dr. Shuo Li, as the principle investigator and primary 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. Dr. Terry Peters, as the co-supervisor, provided guidance and was responsible for the approval of the second manuscript. For each manuscript in this thesis, all other co-authors approved the final draft of the manuscript and their specific contributions are detailed below.

Chapter 2 is an original research article entitled, "Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector." This manuscript was published in the *SPIE Proceedings Volume 10578, Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*. This manuscript was co-authored by Clara M. Tam, Xiaofeng Yang, Sibo Tian, Xi Jiang, Jonathan J. Beitler, and S. Li. Xiaofeng Yang, Sibo Tian, Xi Jiang, and Jonathan J. Beitler assisted with data acquisition and validation of the labelled data.

Chapter 3 is an original research article entitled, "Holistic multitask regression network for multiapplication shape regression segmentation." This manuscript has been submitted to the journal *Medical Image Analysis* in December 2019. This manuscript was co-authored by Clara M. Tam, Dong Zhang, Terry Peters, and Shuo Li. Dong Zhang provided technical advice and assistance.

Acknowledgements

I would like to thank my supervisors, Drs. Shuo Li and Terry Peters for their support and guidance in my endeavours. I am thankful for the many opportunities and challenges that was given to me. It has helped me to gain a lot of invaluable experience and continue to push me beyond my limits. I have truly learned more than I had ever hoped to since first working in this research group. For this, I am truly grateful. I also thank you for your professional mentorship, and faith in my ability to persevere during the many research challenges I experienced.

I am thankful to have Drs. James Lacefield, Robert Bartha, and Terry Thompson 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 essential 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 build professional relationships, acquire a wealth of knowledge, and develop as a researcher. To Dong Zhang, thank you for being a great friend and colleague. Thank you for being the voice of reason in my research and for offering me clarity when it seems like I have hit a road block. Thank you for being there to lend an ear to vent my frustrations, and for imparting your knowledge to me. I am grateful to you for spending the extra time to aid me with technical advice. To Dr. Wufeng Xue, Dr. Xiaoxu He, Dr. Qing Liu, Dr. Chenchu Xu, Dr. Shumao Pang, and Hongbo Wu, thank you for sharing and imparting me with your knowledge and expertise. The skillset and foundations you all have helped me to acquire will be invaluable moving forward.

Most importantly, I would like to thank Dr. Andrew (W.H.) Tse, my family, and my friends. Your patience, support, and encouragement have been a significant part to my success. Andrew, thank you for being a pillar of support during the rough times in my journey. You have been a vital part in all my accomplishments providing me with guidance, encouragement, always inspiring me to do my best, and to have the courage to tackle anything that comes my way. To my parents Damon and Dora, thank you for always supporting me in everything I strive to do, and for your endless encouragement. To Bell and Jude, thank you for being so supportive and caring towards myself and my family. Thank you for all that you have done to help provide me with the opportunity to focus on my goals without worry. To my kendo family and mentors, thank you for the opportunities you have provided me

for self development, and for being a supportive community outside of work.

The computing resources provided through the Southern Ontario Smart Computing Innovation Platform (SOSCIP), 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 SOSCIP platform were performed using the data analytics Cloud at SHARCNET (http://www.sharcnet.ca). I also wish to thank Dr. Jinhui Qin for assisting with the computing environment.

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 Transdisciplinary Bone & Joint Training Award from the Collaborative Training Program in Musculoskeletal Health Research at The University of Western Ontario, Schulich School of Medicine and Dentistry, and The University of Western Ontario.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3D	Three-Dimensional
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CT	Computed Tomography
DSC	Dice Similarity Coefficient
FPN	Feature Pyramid Networks
HaN	Head and Neck
HD	Hausdorff Distance
HMR-Net	Holistic Multitask Regression Network
HOG	Histogram of Oriented Gradients
IoU	Intersection over Union
LOSOCV	Leave-one-subject-out cross-validation
LRN	Linear Regression Network
M^3	Multiplane, Multimodality & Multistructure
mAP	Mean Average Precision
MHz	Megahertz
MRF	Markov Random Fields
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MSVR	Multi-output Support Vector Regression
MTSR	Multitask Shape Regression
N-ResNets	N-Residual Networks
OAR	Organs-at-risk
ReLu	Rectified Linear Unit
ResNet	Residual Neural Network
RF	Radio Frequency
RMSE	Root Mean Square Error
ROI	Regions of Interest
RPN	Region Proposal Network
SRN	Shape Regression Network
SVM	Support Vector Machine
T	Tesla
TE	Time to Echo
TR	Repetition Time

CHAPTER 1

This chapter provides a general introduction to medical image analysis, machine learning, and deep learning concepts. A literature review of current development and challenges of machine learning in medical image segmentation is included in this chapter. The motivation and objectives of applying the benefits of machine learning algorithms will be presented.

1 INTRODUCTION

1.1 Overview

Medical image analysis is a core field of innovation in medical imaging, which is characterized by the use of medical computer vision to analyze the intricacies of the human body [1]. The arrival of digital images in the 1970s [2] and the integration of medical image acquisition devices into the clinical workflow, such as computed tomography (CT) and magnetic resonance imaging (MRI), provided the medical field with great advancements. The digital acquisition of images has provided us a digital lens into the body. This digital vision for healthcare has helped to improve health communication among practitioners, public health researchers, and to patients. Imaging itself has enabled for non-invasive visualization, less invasive surgical methods (e.g. keyhole surgeries [3]), and more precise assessments into a patient's condition. In addition, this has paved the way for more advanced and innovative technology to interpret medical images.

The task of image analysis centres on the importance of the captured semantic information within images. Advancements in this field are centred on the following processing steps for abstract interpretation and quantitative measurements that image analysis embodies [4]:

1) feature extraction – the selection and extraction of distinctive properties or characteristics within the input data for other subsequent steps;
2) segmentation – the process of separating regions of interest from the background and from each other;
3) classification – the process of sorting information into categories according to shared qualities or characteristics;
4) interpretation – determining the conclusions, significance, and implications from all the collected information;
5) measurement – obtaining quantitative values.

1.2 Medical Images

The main types of images focused on in this thesis are CT and MRI. They are also referred to as cross-sectional slice images as they correspond to how an object would appear if it were sliced open along a plane. Both types of images provide detailed information on the anatomy and soft tissues inside the body; however, the details vary from one another based on the acquisition techniques. In some medical cases both CT and MRI images are obtained due to the complementary information they can provide.

1.2.1 Computed tomography (CT))

CT images are reconstructed from the absorption of x-rays. CT uses narrow beams of x-rays, aimed at a patient, quickly rotating around the body to produce a computerized signal. This signal is processed by a computer linked to the x-ray machine to construct a series of cross-sectional image slices of the patient's body [5]. The patient table moves simultaneously through the rotating gantry (ring) housing the x-ray tubes and detectors (see Fig 1.1). Large volumes of data can be acquired quickly and accurately while the patient remains in one position.

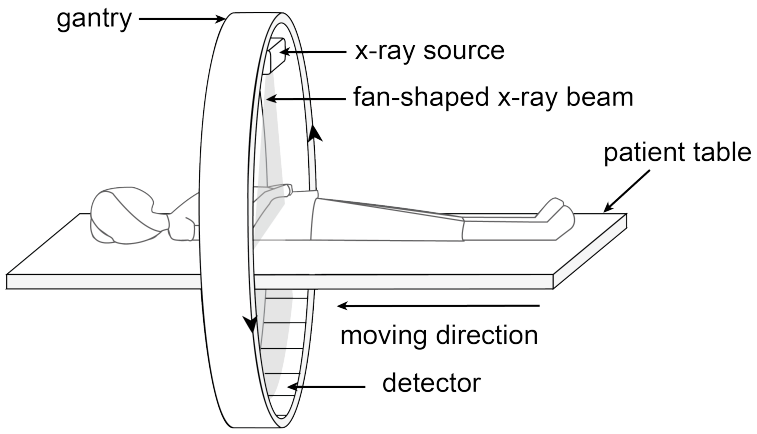


Figure 1.1. A schematic diagram of computed tomography (CT). A CT image slice is reconstructed from the absorption of multiple x-ray projections from different angles. The rotating gantry and patient table move simultaneously to acquire slices of CT images. A sequence of CT images is generated by repeating the image acquisition procedure.

The image data is presented as grey levels based on the level of absorption or attenuation of the ionizing radiation. The linear attenuation coefficient represents the amount of radiation lost by a given thickness of an absorbing material [6, 7]. This increases with the atomic

number and density of the material. This difference in the linear attenuation coefficient between tissues gives rise to contrast in radiographic images. Tissues with low attenuation (e.g. air) will appear dark since very little radiation was absorbed, allowing for most of the radiation to pass through to the detector. Tissues with high attenuation (e.g. bone) will appear bright in the image, as most of the radiation was absorbed, allowing only for a small portion to pass through [5].

During interpretation, the left side of the image is the right side of the patient's anatomy and vice versa. CT images can be displayed individually or stacked together forming a volume or three-dimensional (3D) representation. When viewing a CT volume, the individual volume elements are referred to as voxels as opposed to pixels (picture elements) in a 2D image slice [2].

1.2.2 Magnetic resonance imaging (MRI)

MRI images are generated from the exchange of radio frequency (RF) energy between a patient's body and the imaging system. This is possible due to the intrinsic magnetic properties of the human body; specifically, from hydrogen atoms (protons) which are most abundant within tissues. A proton possesses an innate spin angular momentum rotating about its axis at a constant rate [8–12]. When a strong static external magnetic field is applied, the protons' rotational axes will align with the applied field. Since protons possess an angular momentum, they will precess about the applied field or rotate perpendicularly [13].

The proton's precession rate is directly proportional to the strength of the magnetic field (see Fig. 1.2), 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 magnetic field in tesla (T), and γ is the nucleus-specific gyromagnetic ratio expressed in radian per second per tesla (rad·s^{-1}T^{-1}) or MHz/T for $\frac{\gamma}{2\pi}$ [8, 9].

During the image acquisition procedure, the technician will introduce an RF pulse to disrupt the protons forcing them into a 90° or 180° alignment with the static magnetic field. When the protons' precessing rate or frequency matches the frequency of the applied RF pulse, the protons will begin to resonate and absorb some of the energy from the RF pulse,

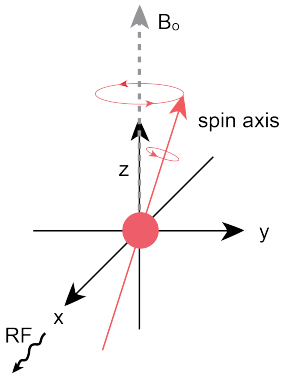


Figure 1.2. Precession of protons in a static magnetic field.

placing them in an excited state. While in the excited state protons will have an increase in electromagnetic energy. When the RF pulse is removed or turned off, the protons will realign with the magnetic field and release the excess electromagnetic energy along the way. This process is known as relaxation. The transmitted RF energy is then picked up by a receiver in the MRI scanner to form the MR images.

Since tissues in the body contain varying amounts of water, MRI detects the electromagnetic fields of the nuclei (protons) to distinguish the differences in density and the shape of tissues throughout the body. Based on how quickly the protons release their excess energy after the applied RF pulse is turned off, one can automatically differentiate between the different tissues in the body.

The relaxation rate (or time) is the most significant factor in producing contrast among distinct types of tissues in an image. Additionally, the intensity of the RF signal is most significant in determining image quality.

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. In other words, the time it takes for the longitudinal signal to regain 63% of its magnetization value. Whereas, T2 is the relaxation time-constant for excited protons to lose phase coherence with each other [9, 10, 14]. In other words, the time it takes for the transverse magnetization to decay to 37% of its original value [15].

MRI acquire image sequences by varying the RF pulses, based on the T1 and T2 relaxation to enhance the quality of select organs. The repetition time (TR) is the time between suc-

cessive pulse sequences applied to the same image slice. 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.

MRI sequences rely on the intrinsic T1 and T2 relaxation properties to generate T1-weighted and T2-weighted scans. T1-weighted sequences have short TR and TE times, whereas T2-weighted have longer TR and TE times. T1 images map the proton density within fatty tissues of the body, such as the bone marrow of the vertebral bodies. Since the cerebral spinal fluid contains no fat it appears dark in T1-weighted images. Conversely, T2 images map the proton density of fatty and water-based tissues. Thus, in T2-weighted images the cerebral spinal fluid would appear bright.

1.3 Feature Extraction

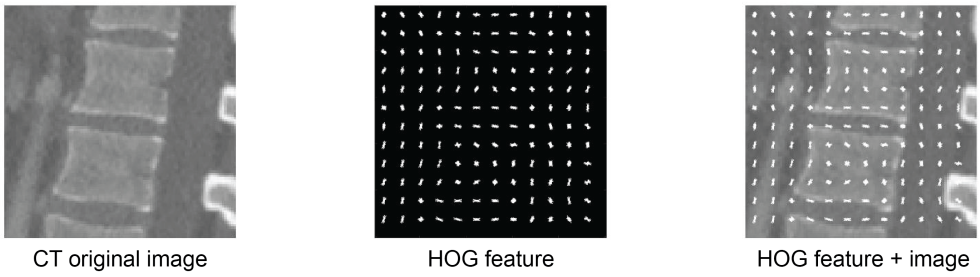


Figure 1.3. Feature extraction using the Histograms of oriented gradients(HOG) feature descriptor. The middle image displays the extracted image features and the image on the far right shows the extracted features overlaid on the original image.

The first processing step, feature extraction, is the most important aspect in computer-based image analysis. Different feature extraction techniques will transform the original data into various forms of representations. These representations can, for example, take the form of gradient orientations [16], a collection of edges [17, 18] contours [19, 20], keypoints [21], texture mapping [22, 23], or skeletons [24, 25]. Some feature extraction techniques will gather enough features capable of reconstructing the input image [26–29]. While other methods will capture invariant properties (e.g. rotation, flips, translation, and scaling) [30], which are considered expected distortions [31]. Fig. 1.3 illustrates an example of feature extraction with a gradient feature descriptor. The image in the centre displays the extracted image features that would be used by a machine learning algorithm to perform classification or regression.

1.4 Segmentation

Segmentation plays a significant role in medical image analysis as a preprocessing step to acquiring valuable quantitative data for medical interpretation. It has allowed for considerable information regarding the anatomy of target structures to be obtainable in an non-invasive manner. The tracking of disease progression [32, 33], localizing pathologies [34], biomechanical modelling [35], simulating biological processes [36], and the evaluation for surgical interventions and therapy [37, 38] are all facilitated by the information obtained from segmentation [39].

This processing step partitions an image's pixels into groups that correlate with the objects present in the image. The representation of the image is changed into a binary image that is easier and more meaningful to analyze. This binary image is often referred to as a segmentation mask, which encodes the spatial layout of all the objects existing in its source image. Within the segmentation process are two levels: 1) Semantic segmentation, the process of assigning a categorical label to each pixel in the image. 2) Instance segmentation, the process of distinctly locating and delineating each object of interest in an image.

Image processing tools for segmentation tasks are abundant spanning across all anatomical regions of the body and on a cellular level. Segmentation techniques can be applied for any type of image (natural images, CT, MRI, ultrasound, PET). Common early segmentation practices involved thresholding, edge detection, region-based, and clustering methods.

Thresholding techniques changes a grey level image into a binary image. It is one of the simplest and fastest methods under the assumption that images are formed by regions of varying grey levels [40]. An image's pixel intensities are divided into two parts, where values greater than or equal to a set intensity value (threshold) becomes the foreground and values less than the threshold is the background. Example thresholding methods include global thresholding (setting a single threshold value to divide the entire image into two groups), local adaptive thresholding (dividing an image into sub-images with different threshold values for each region [41, 42]), and Otsu's thresholding (selecting the optimal threshold value to minimize intra-class variance).

Edge detection methods detects any abrupt changes or discontinuities in image intensity

levels. Although edge detection algorithms can provide detailed shape information they are also hampered by fragmented edges, missing edge segments, and even false edges. The edges detected are often disconnected and would have to be joined to complete the segmentation process. Common edge detection algorithms are Sobel [43–46], Prewitt [44–47], Robert [44–46], Canny [45–47], and Laplacian [45, 46].

Region-based methods group adjacent pixels together with similar or identical features into distinct regions. Regions are connected pixels formed based on a predefined criterion (e.g. colour, texture, or intensity) and unlike edge-based segmentation methods are noise resilient. Region-based techniques are characterized into the following categories: region growing, region splitting, and region merging methods [48]. Region growing uses manually selected seed points. A region is grown by examining and joining neighbouring pixels to that region class as long as it satisfies that class' criterion and no edges are detected [49]. Region splitting and merging algorithms are opposite to region growing where the entire image is split into regions in a top-down fashion. These regions are randomly disconnected, but the elements within these isolated regions are more homogenous than the entire image. Regions can be split into sub-regions and then merged to form new regions with regions that are more suitable [50].

Clustering methods seek to discover similarities existing within the data and partitions it into a specific number of groups or clusters. Clustering is an unsupervised learning method, where a clustering algorithm learns from existing information and is not trained in any form. This segmentation method is mostly used when the classes are known. Similar pixels are clustered together based on exploiting intra-class likeliness and reducing inter-class likeliness [50]. The most commonly known clustering technique is k-means clustering. It is a statistical clustering algorithm that produces k-number of object groups depending on an index of likeness and unlikeness between grouped components in the data. A draw back to clustering methods is the dependency on the initial k value selected. The dependency can be mitigated by running k-means several times with different initial values and choosing the best result; however, this is only valid for low k values. For higher k values more advanced k-means algorithms are required to pick the initial k value. Clustering algorithms also have difficultly with outliers and clusters of varying sizes. This would require the clipping of outliers and generalizing k-means.

However, with the arrival and current hype around deep learning technology, machine learning and specifically neural network-based algorithms are becoming more and more

popular. Machine learning-based methods have allowed for the creation of semi-automatic and fully automatic segmentation algorithms, requiring less human intervention.

1.5 Challenges in medical image segmentation

The development of computer-based image analysis techniques has rapidly evolved from manual techniques to the development of semi-auto and fully automated tools. Despite lacking the intricate skills and expertise of a trained professional, automated and semi-automated image analysis algorithms benefit from being fast, precise, repeatable with constant quality, and are objective. They are capable of relieving human operators from the monotonous and fatiguing parts of the interpretation process [2].

However, the appropriate way to incorporate a prior knowledge on the nature and content of images into these algorithms is still an engineering challenge in the research field. This discrepancy between the discrete pixel representation of an image and the cognitive interpretation by a physician is referred to as the "semantic gap" [4]. Hindering factors that prevent the bridging of this gap consist of the heterogeneity of images, unknown delineation of objects, and the robustness and validation of the developed algorithms [4].

In the design aspect of algorithmic solutions, all constraints from its intended application must be adhered to. Otherwise no matter how outstanding a solution is, "a beautiful algorithmic solution can be virtually useless if the constraints from the application are not adhered to [2]." In the pursuit of innovative engineered solutions, researchers must not lose sight of the practicality from an end-user's perspective.

1.6 Machine learning in medical image segmentation

Developments in information technologies have greatly impacted the technological advancements in healthcare, especially with the integration of artificial intelligence; more specifically machine learning. The continued rise and complexity of data means that machine learning algorithms will increasingly be applied in the medical field [51].

Machine learning is a study of algorithms and statistical models used to perform a specific task without explicit instructions [52]. Machine learning algorithms build a mathematical model from a set of data containing inputs and the desired output [53]. There are three main

categories in which machine learning models fall under: supervised learning, unsupervised learning, and reinforcement learning.

1.6.1 Types of learning

Supervised learning: Traditional supervised learning algorithms are trained with data that contain labels designating the desired output for each corresponding input. This type of learning is the most widely used and established method in the field. Two types of supervised learning algorithms are classification and regression. For a classification task, that determines if an object exists in an image, the output would be a prediction of Boolean values true and false. Regression tasks output continuous numerical values within a range, such as a measurement or a price of an object.

In cases where fully labelled data are unattainable, many machine learning researchers employ alternative techniques to bridge the gap between the disparity of having few labelled data samples and a vast number of unlabelled samples. Full manually-labelled datasets are costly and time-consuming to obtain, especially in cases where domain expertise is required and cannot be repurposed for other objectives. Other types of learning under the supervised learning setting have appeared as alternative solutions: semi-supervised learning, weakly supervised learning, and active learning.

Semi-supervised learning algorithms are trained with data that is incomplete or partially labelled. For this method to work, semi-supervised learning must also rely on structural assumptions within the data to account for the unlabelled samples, such as the notion that points that are close in proximity to one another fall under the same label.

Weakly supervised learning algorithms learn from limited labelled data or noisy labels. For weakly supervised learning, the assumption is knowing that the labels may be imprecise, inexact, or inaccurate [54].

Active learning primarily focuses on estimating the most valuable points for the model to be labelled by a professional. It optimizes the method to acquire training labels in terms of efficiency and cost.

Unsupervised learning: Unsupervised learning algorithms learn the desired output from the data without any predefined labels. They learn the inherent structure of the data from discovering specific patterns through repetitive experiences, such as the grouping or clus-

tering of data points. Cluster analysis techniques are used in unsupervised learning to form these groups with shared attributes. Some examples of unsupervised learning algorithms are autoencoders [55], deep belief networks [56], k-means, generative adversarial networks [57].

Reinforcement learning: Lastly, reinforcement learning algorithms are the part in machine learning that handles sequential decision-making [58]. They are trained in the form of positive or negative feedback in a dynamic/interactive environment. These algorithms are penalized when making a wrong decision and rewarded when the decision is correct. The most notable types of reinforcement learning algorithms are ones designed to compete with human experts in games, such as chess or Go. Reinforcement learning algorithms place an emphasis on finding a balance between known (exploitation) and unknown knowledge (exploration). Unlike supervised learning algorithms, they do not rely on heavy supervision from the labelled sample data, but on a learning agent that interacts with an environment. The agent is the component that makes the decisions of what actions to perform and receives either a penalty or reward for its choices. The agent can use any observed information from the environment or internal rules it possesses.

In order to further understand the functionality of machine learning models the core fundamental concepts required to know are: linear regression, logistic regression, and support vector machines.

1.6.2 Linear regression

A linear regression algorithm is the rudimental algorithm in which every machine learning researcher begins with and forms the fundamental base to understand other machine learning algorithms. A regression problem models the relationship between a dependent variable y and one or more independent variables. When there is only one independent variable x it is a linear regression that can be expressed by the following with m being the slope of the line and b representing the y-intercept: $y = mx + b$.

1.6.3 Logistic regression

Logistic regression algorithms model the probability of an event to occur or the existence of an object [59]. Its most basic form models a logistic function of a binary dependent variable. In machine learning the sigmoid logistic function is used. In a classification task, each object in an image will be assigned a probability in a range between 0 and 1. In most

cases the background is defined as zero while the foreground of the target object is given the value of one.

Logistic regression can also be extended into two other types: multinomial and ordinal regression. Multinomial logistic regression is when there are more than two dependent variables. For a classification task, it would be when the categorical output values are more than two, thus changing the binary task to a multiclass problem. If the multiple categories are ordered, then the regression problem can be modelled by ordinal logistic regression. In machine learning this is also known as ranking learning [60].

1.6.4 Support vector machines

Traditional machine learning techniques, such as the support vector machine (SVM), divide the data into regions with a linear boundary called the hyperplane (decision surface) as shown in Fig. 1.4. SVMs separates the two groups of data by maximizing the margins of the hyperplane [61]. The support vectors in a SVM algorithm are the data points closest to the hyperplane and are the most difficult to classify. For nonlinearly separable patterns a kernel (window) function is used to transform the original data to map into a new higher dimensional space, in which the data becomes linearly separable (see Fig. 1.4).

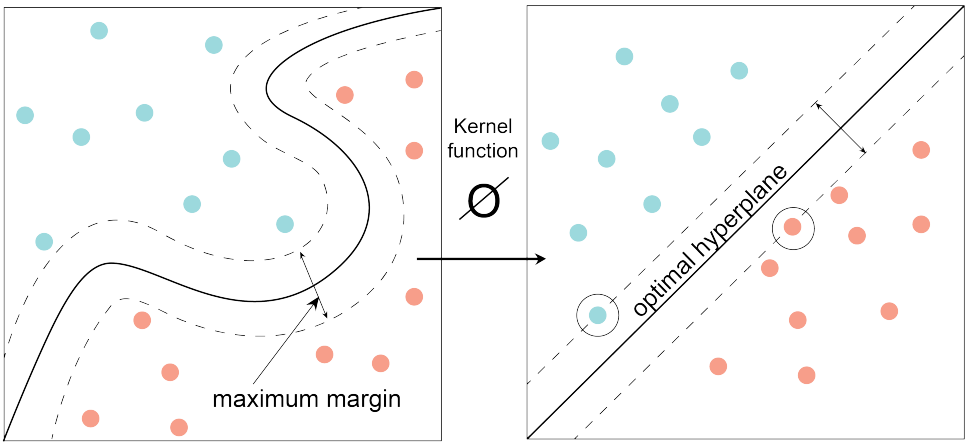


Figure 1.4. The kernel trick. A kernel function transforms the nonlinearly separable data (on the left) to a higher dimensional space (shown on the right), where the data becomes linearly separable. The support vectors are the circled data points in the right image.

The input for a SVM model is a set of training pair samples (the input image and sample features). The output is a set of weights, one for each sample feature whose linear combination predicts the output dependent variable. The weights in a SVM are optimized

to maximize the margin of the hyperplane, thus reducing the number of nonzero weights. Nonzero weights correspond to the support vectors because they "support" the separating hyperplane. Support vectors are unique to the data and are key elements in positioning the hyperplane. If a single support vector is removed the position of the hyperplane will change.

1.7 Deep Learning

The arrival of deep learning algorithms rapidly emerged as the leading technology for medical image analysis [62, 63] based on its ability to map multiple levels of semantic information and its scalability to large amounts of data. This is different from other machine learning methods that reach a plateau in performance. Deep learning algorithms will continue to improve when provided with more data. The larger the dataset the greater the sensitivity the algorithm has to subtle differences.

Deep learning, similar to other machine learning techniques, draws experience from predetermined outcomes based on reference datasets, or learns to formulate and recognize patterns straight from the data itself. Deep learning is a machine learning technique and subfield of machine learning primarily used to classify images through the extraction of visual data. Feature extraction, before the introduction of deep learning algorithms, were all manually crafted and engineered. However, the problem with manually engineered features is determining which ones and for what application they were best suited for. A lot of time and effort would be placed on testing out the several types of feature descriptors and their numerous combinations.

A traditional machine learning algorithm is trained to process and extract manually selected image features (e.g. colour, texture, shape, edge patterns, pixel intensities, and pixel spatial relationships). On the other hand, deep learning automates the feature extraction process, no longer requiring the manually engineered feature extractors prior to training.

Another unique aspect of deep learning is its hierarchical architecture for feature learning. Similar to how the human brain learns to associate distinct characteristics to distinguish a specific object, a deep learning algorithm emulates this behaviour by taking a complex and abstract task, like distinguishing a triangle, and breaks it down into many levels of simpler tasks. The data is essentially represented as a nested hierarchy of concepts.

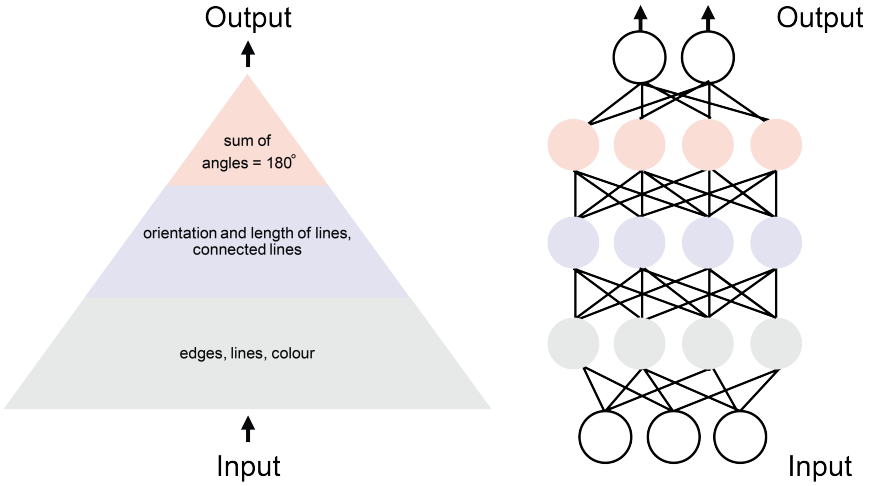


Figure 1.5. Deep learning's hierarchical feature learning. A hierarchy of concepts or characteristics that defines a triangle are presented on the left. This hierarchy pyramid is representative of the levels of extracted information in a deep neural network layer by layer.

For example, the task to determine what is a triangle is a complex and abstract concept. At the lowest level and very first step is to look at the object's lines (a simple concept). How many lines are there? Are there 3 lines and are they connected? Does the sum of their angles equal 180°? With each new level of descriptive information forms a nested hierarchy of concepts (see Fig. 1.5). Inside a deep neural network, these feature levels are represented in the number of layers. The lowest level represents the layers closest to the input layer and the highest level represents the layers closest to the output (see Fig. 1.5). By combining all the individual characteristics, we are able to distinguish a triangle from any other object. In this same manner, a deep neural network combines all the features obtained from each layer into the very last layer before the output. Using the information the algorithm has learned, the trained algorithm can then be applied to new datasets and predict an outcome.

The next few subsections detail the components of a deep neural network and how they function together.

1.7.1 Neural networks

Neural networks, also known as a multilayer perceptron, form the basic architecture to deep neural networks with the term "deep" referring to the network's depth. The concept of the neural network was inspired by the field of neuroscience, whereby an individual node (artificial neuron) in a neural network is representative and mimics aspects of a biological

neuron. A neural network is comprised of a collection of nodes interconnected in various architectures to allow communication between each other. A network's width is the number or nodes or units in a given layer.

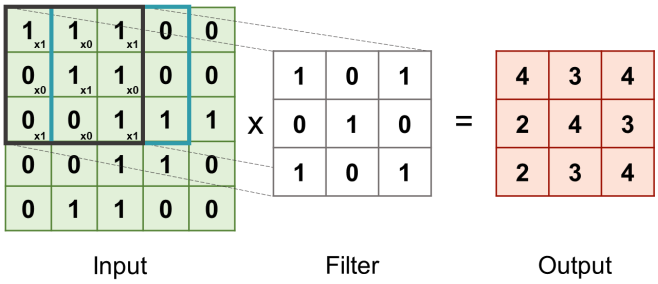


Figure 1.6. Convolution filtering. The convolution filter/kernel performs matrix multiplication with the input image starting from the top left corner. The filter moves across each row of pixels horizontally.

Convolutional neural networks (CNNs): Convolutional Neural Networks (CNNs): CNNs are a type of machine learning algorithm that classifies input data to a category. The architecture of CNNs consists of a sequence of layers. The convolution layer is the main filter component that extracts and generates a predefined number of feature maps computed from the original input image or the output from the previous layer. The convolution filter moves horizontally across each row of pixels in an image starting from the top left corner (see Fig. 1.6). Immediately following the convolution layer is the activation layer. The activation layer is a sigmoid, hyperbolic tangent, or rectified linear units (ReLu) layer that reduces overfitting and exploding gradient effects by introducing nonlinearity to the convolution layer. A pooling layer partitions the previous layer into nonoverlapping rectangles and returns the maximum or average value in each rectangle. Pooling significantly reduces the input size and the number of parameters, thus further preventing the CNN from overfitting. Lastly, the image passes through fully connected or dense layers that generate the network output.

Residual CNNs: Residual CNNs (also known as ResNets) possess a similar structure to regular CNNs; however, with the addition of residual connections and skip connections. These connections enabled deep neural networks to continuously improve with the addition of more layers, creating deeper and deeper networks. ResNets operate in low depth blocks of approximately 20-30 layers that act in parallel [64].

The problem with regular CNNs without the residual and skip connections was that after reaching a certain number of layers there would actually be a decrease in performance. This was due to the vanishing gradient problem when using gradient-based optimization techniques. The vanishing gradient problem was a difficultly found in training artificial neural networks, where the weights would be prevented from updating due to the gradient vanishing by becoming extremely small. The gradient of a network is found using back-propagation, which determines the derivatives of the network layer by layer starting from the final layer all the way to the first layer. A small gradient will prevent the network weights and biases from updating efficiently for each training iteration. The problem arises from mapping a large input space into a small one.

1.7.2 Training methods

Neural networks and deep neural networks can be fast and accurate but are difficult to train. Neural networks are trained in an iterative process through the forward propagation of information and backpropagation, as shown in Fig. 1.7. During forward propagation, the initial information from the input is propagated through each of the hidden units at each layer. Each hidden unit (neuron) performs their transformations on the data received from the previous layer and sends it to the next layer. Once the final layer is reached an output is produced for the corresponding input sample.

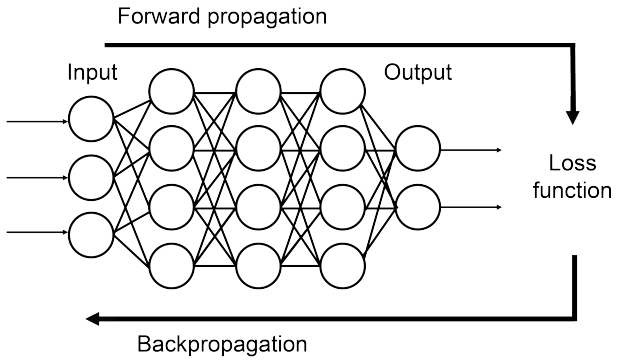


Figure 1.7. Training process of a neural network. Neural networks learn in an iterative process through propagating information forward and backward from an loss function. The network aims to minimize the loss function by updating the network's weights during back-propagation. The new output information is fed back to the loss by forward propagation.

The main goal of the network's training algorithm is to minimize the error of the network's loss function (see Section 1.7.3). Backpropagation allows the flow of information back from the loss function. Information passed to the loss function is propagated and iterated

over the hidden units in a reverse topological order to calculate the final node output. During the iterative process, the weights for each hidden unit are adjusted based on reducing the loss function's error. The two most common strategies to train a network are from scratch or by transfer learning.

From scratch: The initial approach to training a machine learning or deep learning algorithm is to tune the entire network (from scratch/nothing). Training from scratch would require at least over 1000 images of data with the use of data augmentation in order to train a deep learning model. Other than the data itself, a neural network trained from scratch requires its layer weights and biases to be properly initialized. Without proper initialization of a network's weights, many networks would fail to learn anything or return a less than optimal result (getting stuck in a local optima). Careful initialization of a network can help speed up the training process.

The traditional standard practice is to initialize the weights of a neural network model to small random numbers and the biases to zero. Weights should never be initialized to zero otherwise the algorithm would fail due to the inability to change or update the weights. Hidden units connected to the same input and with the same activation function must have different initial weight values in order for the learning algorithm to update the weights. If the weights are the same, then the learning algorithm will constantly update the units in the same way.

Transfer learning: Another method for training a deep learning network with limited data is through the use of transfer learning. Utilizing pre-trained networks on a similar task to initialize the training for a new model has allowed deep neural networks to be trained quicker, be more accurate, and require less training data. Techniques for medical image segmentation are often borrowed from the field of computer vision, which is a scientific field pertaining to how computers can acquire a deep understanding of images or videos on the level of human cognition. This is the case for using pre-trained deep neural networks. Due to the costly nature of accumulating large-scale labelled datasets in the medical field, researchers and scientists have leveraged pre-trained models on a large-scale database for natural images (photography), for example ImageNet. While the end goal may differ, initial low-level image features can be used to facilitate the early training stages of a new deep learning model. Transfer learning has overcome the isolated learning paradigm of conventional machine learning and deep learning algorithms.

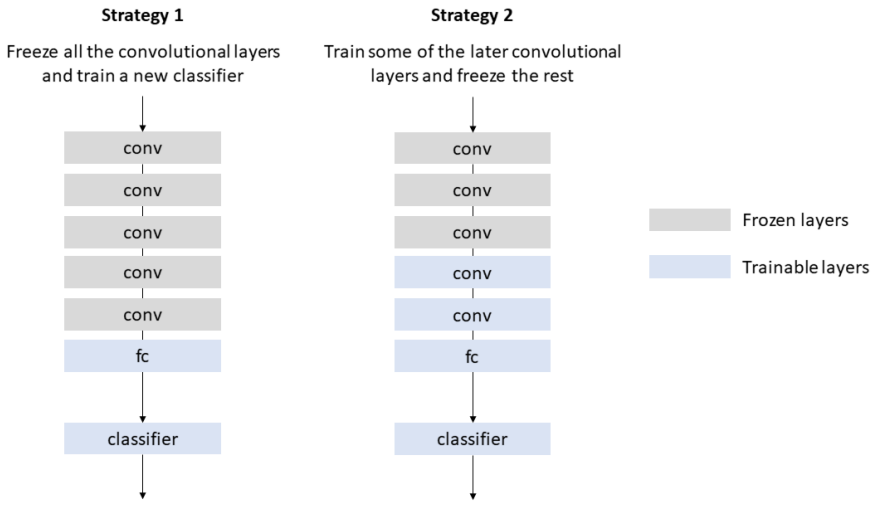


Figure 1.8. Transfer learning strategies.

There are two types of fine-tuning strategies when repurposing a pre-trained model. The first is to freeze all the layers in the pre-trained model and replace the classifier (last layer) with one that is more appropriate for the new task. The second strategy is to freeze a few of the beginning layers in the repurposed model while updating the rest of the later layers (see Fig. 1.8.). The more layers that are frozen represents more information transferred from the previous task. Whereas if your dataset and the previous task differs greatly then only the top few layers could be frozen.

1.7.3 Loss functions

The capacity of a network to approximate the ground truth labels for all the training inputs is quantified by defining a loss function that takes as inputs samples from the training set, weights, and biases. The purpose of the loss function during training is illustrated in Fig. 1.7. During training a network aims at minimizing the loss function to a value as close to zero as possible. If the loss value is high, the loss function penalizes the network by actively changing the weights more frequently. If the loss value is low, the weights will only slightly change since the network is performing well.

Specific loss functions are commonly used for certain tasks as a baseline. For instance, numerical/regression tasks would use the mean squared error as the loss function to calculate the differences between continuous variables. Whereas categorical tasks would use the cross-entropy (log) loss function to calculate the differences between probability distributions [65]. Different tasks possess distinct outputs, and are thus modelled by specific loss functions.

1.7.4 Optimization algorithms

The most efficient way to determine the weights and biases is to use an optimization function, such as stochastic gradient descent. The earliest and main optimization method is stochastic gradient descent. Stochastic gradient descent is a stochastic optimization algorithm that uses and generates random variables. It is an iterative method that optimizes the objective or loss function of a neural network model. The algorithm passes through the training set several times until it converges. After each pass the training set is randomly shuffled.

Other optimization algorithms (e.g. Adam [66] and Adadelta [67]) rely on more advanced techniques, such as momentum and adaptive learning rates. These techniques allow them to converge faster and are easier to implement hyperparameter tuning algorithms like grid search or random search. However, they require more processing time and memory consumption.

1.7.5 Regularization approaches

During training, the accuracy of deep neural networks can continuously converge to perfection but degrade during validation. This is due to the network having memorized the data too closely including the noise. This phenomenon is known as overfitting. Regularization techniques are used in order to prevent a neural network model from overfitting and improve its generalizability.

The most common types of regularization are L1 and L2. Both methods add a regularization term to the model's loss function. L1 (lasso regression) penalizes the absolute value of the weights. Whereas, L2 (rigid regression), also known as weight decay, forces the weights to decay towards zero. Other regularization techniques include data augmentation, dropout, early stopping, and batch normalization.

Data augmentation: Data augmentation is method used to increase the size of the training samples so as to prevent the model from memorizing every variation of the data. Invariant properties or expected distortions of the data can be introduced as augmented samples, such as flips, rotations, scaling, or intensity changes.

Dropout: The idea behind dropout is to randomly drop a certain percentage of neurons in each layer during training by setting the dropped neurons' activation to zero. During

testing or validation all neurons will be active. By dropping some neurons during the training phase, different variations of the model can be obtained. Dropout helps to reduce the interdependent learning between the nodes.

Early Stopping: A network's training can be stopped before it begins to overfit. Early stopping is commonly used in practice and is implemented by measuring the accuracy or loss on an isolated test set. When the test performance ceases to improve the training is stopped.

Batch Normalization: Batch normalization is a method to improve the performance, speed, and stability of trained neural network models [68]. To achieve batch normalization, a normalization step is performed to fix each layer's inputs means and variances for each mini-batch during training. Similar to dropout, batch normalization forces each layer in a network to be robust to variation in its input. Each hidden unit is multiplied by the standard deviation and subtracts the mean of the mini-batch at each step of training. Both the standard deviation and mean randomly fluctuates due to the random samples chosen in each mini-batch being different at each step.

1.8 Clinical relevance

Advanced semi-automatic or fully automatic image segmentation techniques have yet to fully make its way into the clinical domain despite several research and scientific studies conducted on them. Complex pipelines and algorithms that are computationally heavy are a few concerns regarding to the practical use of some of these machine learning design implementations. Ensuring the validity and reliability of new emerging algorithms is yet another concern within the machine learning research field [69, 70].

All the while, advancements in new imaging technology have further increased the amount of imaging exams and thus increased the workload of physicians. Radiologists perform thousands of exams per year and would be required to exam a single image every 3 to 4 seconds to keep up with demand [71]. Automation can improve the services provided by healthcare systems with algorithms for image processing and analysis. The potential and goal of automated tools serve to alleviate menial tasks and deficiencies in technical knowledge enabling physicians to focus more time and effort into the diagnostic and therapeutic aspects of patient care [69].

However, single task or patient specific segmentation models [72] poses the challenge of reusability prompting for more general and holistic methods transferable to other applications. Research conducted for specific applicational use is still of immense value and serve important purposes, but in a grander perspective this holds less impact than solutions catering to generalized or multiple applications. In the perspective of advancing the current state of technology, physicians place a higher importance on and are looking to researchers and engineers to work towards addressing the concerns and hurdles preventing machine learning solutions from transitioning from scientific research and development to commercial and clinical use [69, 73]. One such hurdle is generalizing automated segmentation solutions to deliver a more standardized approach to the task.

1.9 Thesis objectives

The overarching theme of this thesis is to leverage the advancements of machine learning and deep learning technology to develop computer-aided tools appropriate for multiapplication medical imaging analysis. The expectation is to facilitate and aid in diagnostic, prognostic, and treatment planning practices to improve clinical workflow. The chapter specific objectives of this thesis are introduced below.

In Chapter 2, the objective was to perform automated segmentation for organs-at-risk in head and neck CT images comparable or superior to state-of-the-art methods by formulating the segmentation task as a boundary regression problem. The goal was to assess the flexibility and feasibility of the boundary regression approach to adapt to a wide range of organs with various shapes.

In Chapter 3, the objective was to develop a multiapplication framework using a holistic multitask regression approach to tackle the limitations in medical image analysis. The use of multitask learning was hypothesized to enable our proposed algorithm to generalize to more tasks.

In Chapter 4, an overview and a summary of the important findings and conclusions in Chapters 2 and 3 will be presented. The limitations and on-going challenges of current technology will be discussed. The thesis is concluded with potential ideas to explore or develop in future studies based on this research.

References

[1] T. M. D. (ne Lehmann), H. Handels, K. H. M.-H. (ne Fritzsche), S. Mersmann, C. Palm, T. Tolxdorff, *et al.*, "Viewpoints on medical image processing: from science to application," *Current Medical Imaging Reviews*, vol. 9, no. 2, pp. 79–88, 2013.

[2] K. D. Toennies, *Guide to Medical Image Analysis: Methods and Algorithms*. Springer Publishing Company, Incorporated, 2nd ed., 2017.

[3] N. Goel, A. Yadav, and B. M. Singh, "Medical image processing: A review," *2nd IEEE International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity, CIPECH 2016*, pp. 57–62, 2017.

[4] T. M. Deserno, *Fundamentals of Medical Image Processing*, pp. 1139–1165. Springer Berlin Heidelberg, 2011.

[5] T. Osborne, C. Tang, K. Sabarwal, and V. Prakash, "How to interpret an unenhanced CT brain scan. Part 1: Basic principles of computed tomography and relevant neuroanatomy," *South Sudan Medical Journal*, vol. 9, no. 3, pp. 67–69, 2016.

[6] J. T. Hathcock and R. L. Stickle, "Principles and concepts of computed tomography," *The Veterinary clinics of North America. Small animal practice*, vol. 23, no. 2, pp. 399–415, 1993.

[7] M. H. McKetty, "The AAPM/RSNA physics tutorial for residents: X-ray attenuation," *Radiographics*, vol. 18, no. 1, pp. 151–163, 1998.

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

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

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

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

[12] S. W. Atlas, *Magnetic Resonance Imaging of the Brain and Spine*, vol. 1 of *LWW medical book collection*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2009.

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

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

[15] 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: a review publication of the Radiological Society of North America, Inc*, vol. 29, no. 5, pp. 1433–1449, 2009.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 1, pp. 886–893, 2005.

[17] N. Prajapati, A. Kumar Nandanwar, and G. S. Prajapati, "Edge histogram descriptor, geometric moment and sobel edge detector combined features based object recognition and retrieval system," *International Journal of Computer Science and Information Technologies*, vol. 7, no. 1, pp. 407–412, 2016.

[18] K. Mikolajczyk, A. Zisserman, and C. Schmid, "Shape recognition with edge-based features," pp. 79.1–79.10, 2012.

[19] Y. Xu, Y. Quan, Z. Zhang, H. Ji, C. Fermuller, M. Nishigaki, *et al.*, "Contour-based recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3402–3409, 2012.

[20] D. You, S. Antani, D. Demner-Fushman, and G. R. Thoma, "A contour-based shape descriptor for biomedical image classification and retrieval," *Document Recognition and Retrieval XXI*, vol. 9021, p. 90210, 2013.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] P. Banerjee, A. K. Bhunia, A. Bhattacharyya, P. P. Roy, and S. Murala, "Local neighborhood intensity pattern–A new texture feature descriptor for image retrieval," *Expert Systems with Applications*, vol. 113, pp. 100–115, 2018.

[23] A. Humeau-Heurtier, "Texture feature extraction methods: A survey," *IEEE Access*, vol. 7, pp. 8975–9000, 2019.

[24] J. Ding, Y. Wang, and L. Yu, "Extraction of human body skeleton based on silhouette images," in *2010 Second International Workshop on Education Technology and Computer Science*, vol. 1, pp. 71–74, 2010.

[25] Z. Yang, F. Guo, and P. Dong, "Skeleton extraction of a specified object in the gray image based on geometric features," in *Information Computing and Applications* (C. Liu, L. Wang, and A. Yang, eds.), (Berlin, Heidelberg), pp. 161–168, Springer Berlin Heidelberg, 2012.

[26] F. P. Kuhl and C. R. Giardina, "Elliptic fourier features of a closed contour," *Computer Graphics and Image Processing*, vol. 18, no. 3, pp. 236–258, 1982.

[27] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489–497, 1990.

[28] Z. Yu, T. Li, N. Yu, Y. Pan, H. Chen, and B. Liu, "Reconstruction of hidden representation for robust feature extraction," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–24, 2019.

[29] Z. Bahaoui, R. Benouini, H. EL Fadili, K. Zenkouar, and A. Zarghili, "Comparative study of exact continuous orthogonal moments applications: Local feature extraction and data compression," *Transactions on Machine Learning and Artificial Intelligence*, vol. 5, no. 4, 2017.

[30] S. Urooj and S. P. Singh, "Geometric invariant feature extraction of medical images using Hu's invariants," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1560–1562, 2016.

[31] Ø. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition-A survey," *Pattern Recognition*, vol. 29, no. 4, pp. 641–662, 1996.

[32] V. Grau, A. U. J. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–458, 2004.

[33] H. Greenspan, A. Ruf, and J. Goldberger, "Constrained gaussian mixture model framework for automatic segmentation of MR brain images," *IEEE Transactions on Medical Imaging*, vol. 25, no. 9, pp. 1233–1245, 2006.

[34] A. El-Baz, A. Farag, G. Gimel'farb, R. Falk, M. Abou-El-Ghar, and T. El-Diasty, "A framework for automatic segmentation of lung nodules from low dose chest CT scans," vol. 3, pp. 611–614, 2006.

[35] S. Iyer, B. A. Christiansen, B. J. Roberts, M. J. Valentine, R. K. Manoharan, and M. L. Bouxsein, "A biomechanical model for estimating loads on thoracic and lumbar vertebrae," *Clinical Biomechanics*, vol. 25, no. 9, pp. 853–858, 2017.

[36] M. Prastawa, E. Bullitt, and G. Gerig, "Simulation of brain tumors in MR images for evaluation of segmentation efficacy," *Medical Image Analysis*, vol. 13, no. 2, pp. 297–311, 2009.

[37] O. Ecabert, J. Peters, H. Schramm, C. Lorenz, J. von Berg, M. J. Walker, *et al.*, "Automatic model-based segmentation of the heart in CT images," *IEEE Transactions on Medical Imaging*, vol. 27, no. 9, pp. 1189–1201, 2008.

[38] A. El-Baz, A. A. Farag, S. E. Yuksel, M. E. A. El-Ghar, T. A. Eldiasty, and M. A. Ghoneim, *Application of Deformable Models for the Detection of Acute Renal Rejection*, pp. 293–333. New York, NY: Springer New York, 2007.

[39] A. Elnakib, G. Gimel'farb, J. Suri, and A. El-Baz, *Medical Image Segmentation: A Brief Survey*, pp. 1–39. 2011.

[40] A. Norouzi, M. Rahim, A. Altameem, T. Saba, A. Ehsani Rad, A. Rehman, *et al.*, "Medical image segmentation methods, algorithms, and applications," *IETE Technical Review*, vol. 31, pp. 199–213, 2014.

[41] C.-H. Teng, W.-H. Hsu, T. A. Busey, B. L. Schneider, K. Moses, W.-S. Zheng, *et al.*, *Local adaptive thresholding*, pp. 939–939. Boston, MA: Springer US, 2009.

[42] S. Bala and A. Kumar, "A brief review of image segmentation techniques," *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)*, vol. 5, no. 5, 2016.

[43] O. Vincent and O. Folorunso, "A descriptive algorithm for sobel image edge detection," 2009.

[44] R. K. Singh, S. Shekhar, R. B. Singh, and V. Chauhan, "A comparative study of edge detection techniques," *International Journal of Computer Applications*, vol. 100, no. 19, pp. 5–8, 2014.

[45] 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 Graphics & Vision*, vol. 12, no. 13, pp. 29–32, 2012.

[46] M. Ansari, D. Kurchaniya, and M. Dixit, "A comprehensive analysis of image edge detection techniques," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 12, pp. 1–12, 2017.

[47] D. Dey and D. Polley, "Edge detection by using canny and prewitt," *International Journal of Scientific & Engineering Research*, vol. 7, no. 4, pp. 251–254, 2016.

[48] 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.

[49] M. Kaur and P. Goyal, "A review on region based segmentation," *International Journal of Science and Research*, vol. 4, no. 4, pp. 2319–7064, 2013.

[50] M. S. Sonawane and C. A. Dhawale, "A brief survey on image segmentation methods," *International Journal of Computer Applications*, vol. 975, pp. 8887–8892, 2015.

[51] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, 2019.

[52] D. Cielen, A. D. Meysman, and M. Ali, *Introducing Data Science*. Introducing Data Science, Manning Publications, 2016.

[53] C. M. Bishop, *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer, 2006.

[54] Z. H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[55] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," *ICML Unsupervised and Transfer Learning*, pp. 37–50, 2012.

[56] Y. H. Hua, J. Guo, and H. Zhao, "Deep belief networks and deep learning," in *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things*, pp. 1–4, 2015.

[57] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al.*, "Generative adversarial networks," 2014.

[58] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[59] J. Tolles and W. J. Meurer, "Logistic regression: Relating patient characteristics to outcomes," *JAMA - Journal of the American Medical Association*, vol. 316, no. 5, pp. 533–534, 2016.

[60] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," *Advances in Neural Information Processing Systems*, 2003.

[61] X. Zhang, *Support Vector Machines*, pp. 941–946. Boston, MA: Springer US, 2010.

[62] M. I. Razzak, S. Naz, and A. Zaib, *Deep Learning for Medical Image Processing: Overview, Challenges and the Future*, pp. 323–350. Springer International Publishing, 2018.

[63] D. Shen, G. Wu, and H.-i. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[64] A. Veit, M. J. Wilber, and S. J. Belongie, "Residual networks are exponential ensembles of relatively shallow networks," *CoRR*, vol. abs/1605.0, 2016.

[65] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *Schedae Informaticae*, vol. 25, pp. 49–59, 2016.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[67] M. D. Zeiler, "Adadelta: An adaptive learning rate method," 2012.

[68] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep detwork training by reducing internal covariate shift," 2015.

[69] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine learning and deep learning in medical imaging: Intelligent imaging," *Journal of Medical Imaging and Radiation Sciences*, vol. 50, no. 4, pp. 477–487, 2019.

[70] Z. Zhang and E. Sejdić, "Radiological images and machine learning: Trends, perspectives, and prospects," *Computers in Biology and Medicine*, vol. 108, pp. 354–370, 2019.

[71] R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, *et al.*, "The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload," *Academic Radiology*, vol. 22, no. 9, pp. 1191–1198, 2015.

[72] J. Thirumaran and S. Shylaja, "Medical image processing-an introduction," *International Journal of Science and Research (IJSR)*, vol. 4, pp. 1197–1199, 2014.

[73] C. W. Ho, D. Soon, K. Caals, and J. Kapur, "Governance of automated image analysis and artificial intelligence analytics in healthcare," 2019.

CHAPTER 2

The contents of this chapter was previously published in the SPIE Proceedings Volume 10578, Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging: C. M. Tam, X. Yang, S. Tian, X. Jiang, J. J. Beitler, and S. Li "Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression", Proc. SPIE 10578, Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, 1057824 (12 March 2018); https://doi.org/10.1117/12.2292556

2 AUTOMATED DELINEATION OF ORGANS-AT-RISK IN HEAD AND NECK CT IMAGES USING MULTI-OUTPUT SUPPORT VECTOR

2.1 Introduction

Head and neck (HaN) cancers develop in more than 550,000 people with a mortality rate of approximately 300,000 fatalities occurring annually worldwide [1]. Anatomical delineation of critical structures is a pre-processing step for diagnosing, treatment planning, and radiation therapy of HaN cancers in order to minimize administered radiation dosage. Accurate segmentation is crucial to ensure correct diagnosis and treatment of the affected regions. Current manual segmentation is limited to physician training and experience. In comparison, machine learning algorithms can utilize a repository of samples from various physicians to standardize segmentation practices and enhance the quality of results [2]. This integrated technology provides reduced laborious segmentation tasks and decreased diagnosis and prognosis time [3]. Although semi-automated and fully automated segmentation algorithms have been developed and are available for delineating organs-at-risk (OAR), their applications are limited for many structures where two-dimensional manual contouring in the axial plane is still common practice [4].

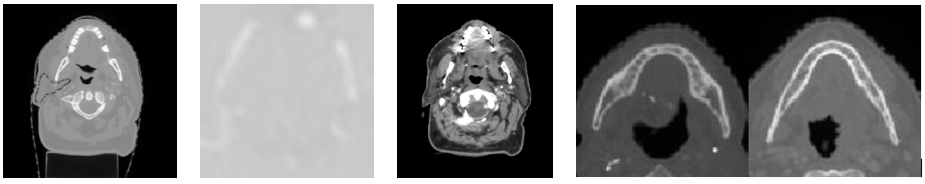
2.1.1 Related works

Ibragimov et al. [5] proposed a four step deep learning method using convolutional neural networks (CNN) and Markov random fields (MRF) to identify and segment the OAR in HaN CT images. The four step method included voxel classification using CNN, smoothing with MRF, removal of small components, and dilate-erode operations to achieve the final segmentation results. Although the CNN coupled with MRF method [5] implemented

state-of-the-art deep learning, their results illustrated a huge gap in dice coefficients among the various OAR from 37.4% [5] for the chiasm to 89.5% [5] for the mandible. This demonstrates the limitation for organs with much lower contrast. Other methods like atlas-based registration coupled with machine learning was implemented at the head and neck auto-segmentation challenge for parotid glands [4], brainstem and mandible [6] segmentation. However, the results for soft tissue organs, such as the brainstem still proved challenging for automated algorithms requiring user correction of approximately half of the slices [4, 6]. The atlas-based algorithm demonstrated in the 2010 head and neck challenge for parotid glands, did however demonstrated good correlation achieving overlap results of around 83% [4].

2.1.2 Challenges

The limitations of these segmentation algorithms stems from the challenges they are faced with 1) complex and irregular morphology, 2) poor soft tissue discrimination of the OAR in CT images, 3) artefacts from dental fillings, 4) variability of patient's anatomy, and 5) inter-observer variability [4, 5, 7] (See Fig. 2.1).



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

Figure 2.1. Illustration of the challenges faced by automated segmentation algorithms: (a) displays an example of the complex and irregular morphology of the parotid gland, (b) depicts the poor soft tissue contrast of the brainstem, (c) dental artefacts, and (d) variability among patient anatomy.

In this study, we propose a state-of-the-art model for the automated segmentation of OAR using a multi-output support vector regression (MSVR) [8, 9] machine learning algorithm for automated delineation of OAR in HaN CT images. Leveraging a supervised machine learning algorithm in a holistic regression fashion [8, 9] the proposed model is capable of performing robust, accurate, and efficient predictions. Holistic regression refers to having:

1) An output where the locations of all the boundary points are simultaneously regressed. In doing so, the global shape guides the points previously learnt from the training data.

2) An input where each boundary point is regressed using the full image as a signal, enabling the boundary regression model to capture the full context of a boundary point.

2.2 Multi-output support vector regression (MSVR)

The proposed framework consists of two processes: 1) the training phase to learn the nonlinear regression model (Eq. 2.1) from the training dataset; and 2) the testing phase that directly predicts the boundaries of the OAR for the input images from the learnt model. The segmentation model learns a nonlinear mapping function F which is assigned to each image, given by the vector X and a vector Y of P target values ($P = 2p$). This mapping function is expressed as

$$Y = F(X), \tag{2.1}$$

$$Y = \{y^{(j)} | j = 1 : P\} = \{(locx_1, locy_1), ..., (locx_p, locy_p)\}, \tag{2.2}$$

$$X = f(I) = [X^{HOG}], \tag{2.3}$$

where Y is the boundary coordinate vector corresponding to the object segmentation in image I and X is the image feature vector extracted from the input image I using the histogram of oriented gradients (HOG) [10]. The object boundary points are represented by a set of p points, $y_i = (locx_i, locy_j)|_{j=1:p}$. $(locx_i, locy_j)$ is the jth boundary point's coordinates. The value of p can determine different resolutions of the boundary shape. This mapping function is fulfilled by the multi-output support vector regression (MSVR) [8, 9]. The overall framework of the MSVR is illustrated in Fig. 2.2. The HOG feature descriptor divides the image into small connected cells, and then computes a histogram of gradient directions or edge orientations for the pixels within each cell.

MSVR is a sparse kernel machine capable of modelling highly nonlinear mapping functions [8]. The algorithm's ability to predict all the boundary points in the output vector Y of Eq. 2.1 simultaneously and dependently ensures the capability of the regression segmentation to handle highly diverse boundaries. Using the radial basis function (RBF) kernel MSVR transforms the data into a higher dimensional feature space ϕ to perform linear separation. The RBF kernel is constructed based on feature X^{HOG}.

The objective of the MSVR is to learn a kernel regressor:

$$Y = W\phi(X) + b, \tag{2.4}$$

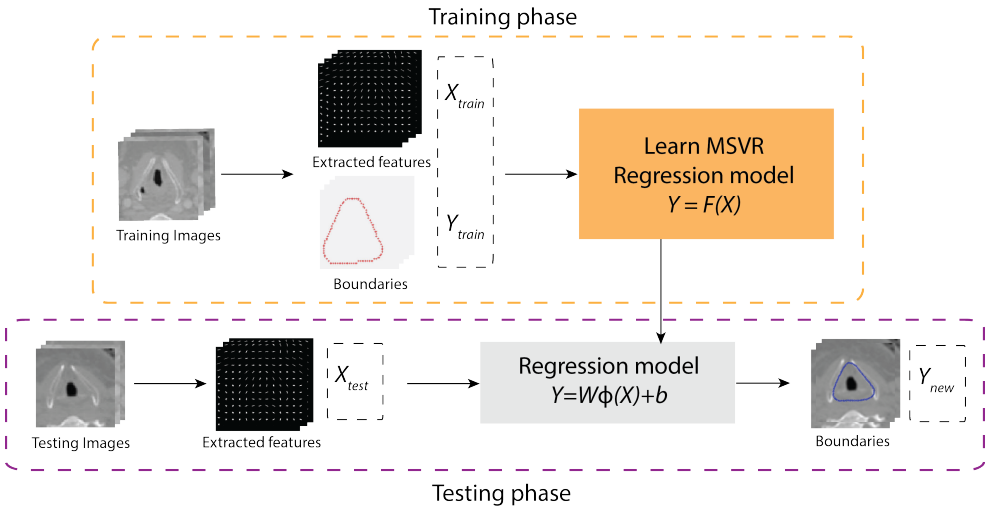


Figure 2.2. A schematic of the overall multi-output support vector regression (MSVR) framework. Our model consists of two phases: 1) training where the regression model learns the mapping function $Y = F(X)$ and 2) testing whereby the model predicts the boundary of new unseen input images.

where $\phi(\cdot) : \mathbb{R}^N \xrightarrow{\phi(\cdot)} \mathbb{R}^H$ is a nonlinear transformation to a higher H-dimensional space, $Y = (y^{(1)}, ..., y^{(P)})$ is the boundary coordinate output, and $X = (x^{(1)}, ..., x^{(N)})$ is the image feature input (N being the number of inputs). Eq. 2.4 denotes for every N dimensional image feature input X composed of individual features, a P-dimensional boundary output Y is simultaneously regressed by a weight vector $W = (w^{(1)}, ..., w^{(P)})$ and the bias parameter $b = (b^{(1)}, ..., b^{(P)})^T$, where T is the transpose notation and $P = 2p$. Transposing a vector transforms a column vector into a row vector or a row vector into a column vector. To provide a holistic regression of all boundary points the goal is to learn the optimal parameters of W and b from a labelled training dataset $\mathcal{D} = \{(X_i, Y_i)|i = 1 : m\}$ with m number of OAR images by solving the optimization problem with MSVR [11, 12]. The optimization expression is denoted as follows:

$$\min_{W,b} L(W, b) = \min_{W,b} \sum_{j=1}^{P} \left(\frac{1}{2} \|w^{(j)}\|^2 + C \sum_{i=1}^{m} \mathcal{L}\Big(Y_i - (\phi(X_i)^T w^{(j)} + b^{(j)})\Big)\right), \tag{2.5}$$

where each training image includes a feature vector of N descriptive variables $X_i \in \mathbb{R}^N$ and boundary points' locations vector with P target variables $Y_i = (y_i^{(1)}, ..., y_i^{(P)}), P = 2p$. The loss term $\mathcal{L}(\cdot)$ is defined as the $\in$ - insensitive loss function. The best solution of W can be expressed as a linear combination of the training samples in the transformed feature space

[11, 12]:

$$w^{(j)} = \sum_{i=1}^{m} \phi(X_i)\beta_i^{(j)}, \quad j = 1 : P. \tag{2.6}$$

Additionally, the best solution of b can be expressed as[9]:

$$b^{(j)} = Y_i^{(j)} - \in -\Big(\sum_{i=1}^{m} \phi(X_i)\beta_i^{(j)} \Big) \cdot \phi(X_{test}), \quad j = 1 : P. \tag{2.7}$$

Once the parameters of W and b are solved, the P-dimensional output Y (boundary coordinate vector) for each new input X_{test} can be computed as:

$$Y_{new} = \sum_{i=1}^{m} \beta_i^T \phi^T(X_i)\phi(X_{test}) = \sum_{i=1}^{m} \beta_i^T K(X_i, X_{test}). \tag{2.8}$$

Where the RBF kernel function is given by:

$$\begin{aligned} K(X_i, X_{test}) &= \phi^T(X_i)\phi(X_{test}), \\ &= \exp(-\gamma \|X_i - X_{test}\|^2), \quad \gamma > 0 \end{aligned} \tag{2.9}$$

which is the dot product between a training sample X_i and the new input vector X_{test} in the feature space ϕ. Here γ represents the term $\frac{1}{2\sigma^2}$, and $\| \cdot \|^2$ denotes the Euclidean norm.

2.2.1 Training phase

1) *ROI selection:* ROIs are automatically cropped and extracted based on two manually selected landmarks. The first landmark is located near the top left corner of the organ of interest and the second landmark is located near the bottom right corner. A square bounding box crops the ROIs with its centre being the middle of the two selected landmarks.

2) *Boundary delineation from ROI images:* The boundary was delineated by an expert physician for each ROI and is used as the training dataset. To smoothly approximate the complex shapes, the number of boundary points for representing each of the OAR in Eq. 2.2 was empirically set to $p = 100$ points. For each boundary, the first point is fixed while the rest $(p - 1)$ are sampled evenly along the outer boundary of the OAR in a clockwise direction. The boundary points' locations are stored in a vector with P target variables (See Eq. 2.2).

3) *Compute image descriptor for each ROI:* Based on the occurrences of gradient orientation in localized portions of an image, HOG [10] captures the main shape of the

anatomical structure in an image. Using HOG, the image descriptor was computed for each ROI to extract the shape features of the OAR. The extracted image features were stored in a feature vector represented by Eq. 2.3.

4) *Using MSVR to learn the mapping function:* The MSVR algorithm (as described in Section 2.2) was used to obtain the optimal parameters W and b in Eq. 2.4. The algorithm was computed in MATLAB R2017a.

2.2.2 Testing phase

For undelineated ROIs its image descriptor X_{test} was computed. Using the learned mapping function (Eq. 2.4), regression segmentation was performed using MSVR [11, 12] to predict the desired boundary Y_{new} for each new input X_{test}. From the regressed boundary, segmentation of the OAR can be directly obtained.

2.3 Experiment Configurations

2.3.1 Dataset

A dataset of 3D CT images consisting of organs in the HaN region were obtained from 56 anonymous patients. All images in the dataset were axially reconstructed and had a pixel width and height of 1.27 mm, voxel depth of 1.25 mm, and a matrix dimension of 512 $\times$ 512. Approximate ground truth delineations were performed by experienced radiologist for the following OAR: brain stem in 56 images, both cochleae in 53 images, esophagus in 50 images, eye globes in 56 images, larynx in 49 images, left lens in 51 images, right lens in 50 images, lips in 46 images, mandible in 32 images, oral cavity in 42 images, both parotid glands in 52 images, spinal cord in 45 images, left submandibular gland in 39 images, right submandibular gland in 40 images, and thyroid in 25 images. The total number of objects segmented was 847.

2.3.2 Training configurations

The proposed model was executed on a single central processing unit (CPU) computer with AMD A10-6700 processor at 3.7 GHz and 12 GB of memory. Preprocessing ROI selection in the training phase took approximately 2.75 minutes per OAR. Training the mapping function for the model along with testing took less than a second per OAR. Giving a total training time of 49.6 minutes for all the OAR.

2.3.3 Evaluation metrics

Two types of metrics were used to evaluate the performance of the segmentation the dice similarity coefficient (DSC) and the leave-one-subject-out cross-validation (LOSOCV) [13, 14]: DSC was used to measure the percentage of overlap between manually segmented boundaries and automatically segmented boundaries of the OAR's anatomical structures. The DSC is given by the following expression

$$DSC = \frac{2|M \cap A|}{|M| + |A|}, \tag{2.10}$$

where M is the area of expert manual segmentation, A is the area of the automated segmentation, and $M \cap A$ is the area of overlap between M and A. DSC has a range of (0,1) where the change from 0 to 1 represents no overlap to complete overlap.

The LOSOCV computes statistics on the left out samples. This metric is a special case of k-fold cross-validation where k is the number of instances in the data. In the case of the LOSOCV, $k = 1$. For each iteration, the model is trained with the entire dataset except for a single observation which is used for testing. The average error or mean squared error (MSE) and root mean square error (RMSE) are computed and used to evaluate the model. MSE tells you how close the regression line is to a set of points. This is accomplished by taking the distances from the data points to the regression line and squaring them. The distances are the "errors". We compute the norm differences and scale them by the number of points. MSE is given by the following expression

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_{new})^2, \tag{2.11}$$

where Y_i is the vector of observations and Y_{new} is the vector of n predictions. The squaring is necessary to remove any negative signs, and also gives more weight to larger differences. This is called the MSE since the process determines the average of a set of errors. The smaller the MSE indicates that the data has low variance from the line of best fit. The RMSE on the other hand refers to the square root of the variance of the residuals and is expressed as $RMSE = \sqrt{MSE}$, where RMSE indicates the absolute fit of the model's predicted values to the observed data points [13]. The RMSE is the distance of a data point from the fitted line (the variance), that is measured along a vertical line.

2.4 Results and Discussion

2.4.1 Overall performance

Segmentation results of the OAR compared against manual segmentation performed by an expert radiologist are illustrated in Fig. 2.3 and 2.4 with their corresponding DSC, MSE, and RMSE results listed in Table 2.1. The approximate ground truth or manual segmentation is delineated with the red asterisks contour and the automated segmentation results are delineated with the blue asterisks contour (See Fig. 2.3 and 2.4). Results varied from 66.9% DSC for the left cochlea to 93.8% DSC for the left eye globe. In addition, the MSE for each OAR came out to be less than 1% with the variance of the predicted data compared to the approximate ground truth data to be from 2.3% for both eye globes to 7% for the left parotid gland (See Table 2.1). The comparison of corresponding organs between our algorithm and a state-of-the-art convolutional neural network-based (CNN) segmentation method is displayed in Table 2.2.

Table 2.1. MSVR's overall performance. DSC is given in terms of mean (± standard deviation). The mean squared error (MSE) and root mean square error (RMSE) results show how close the approximate ground truth boundary points are to the predicted boundary points and the percentage of variance respectively. MSE and RMSE were divided by the sample number to yield the percentage results.

Organs	**DSC (%)**	**MSE (%)**	**RMSE (%)**	**No. of Images**
Brainstem	91.3 ± 4.5	0.16	3.97	56
Cochlea left	66.9 ± 17.2	0.07	2.73	53
Cochlea right	71.5 ± 15.2	0.06	2.45	53
Esophagus	81.6 ± 8.1	0.15	3.91	50
Eye globe left	93.8 ± 2.7	0.05	2.29	56
Eye globe right	93.7 ± 2.9	0.05	2.33	56
Larynx	89.1 ± 5.2	0.17	4.12	49
Lens left	75.0 ± 11.1	0.06	2.56	51
Lens right	74.4 ± 11.7	0.06	2.39	50
Lips	76.7 ± 8.7	0.12	3.44	46
Mandible	85.2 ± 5.3	0.10	3.16	32
Oral Cavity	86.9 ± 7.4	0.34	5.82	42
Parotid left	82.5 ± 7.0	0.49	7.04	52
Parotid right	82.2 ± 8.5	0.40	6.38	52
Spinal cord	83.1 ± 8.5	0.05	2.31	45
Submandibular left	87.2 ± 5.9	0.17	4.09	39
Submandibular right	87.1 ± 5.5	0.22	4.72	40
Thyroid	78.5 ± 6.9	0.39	6.24	25

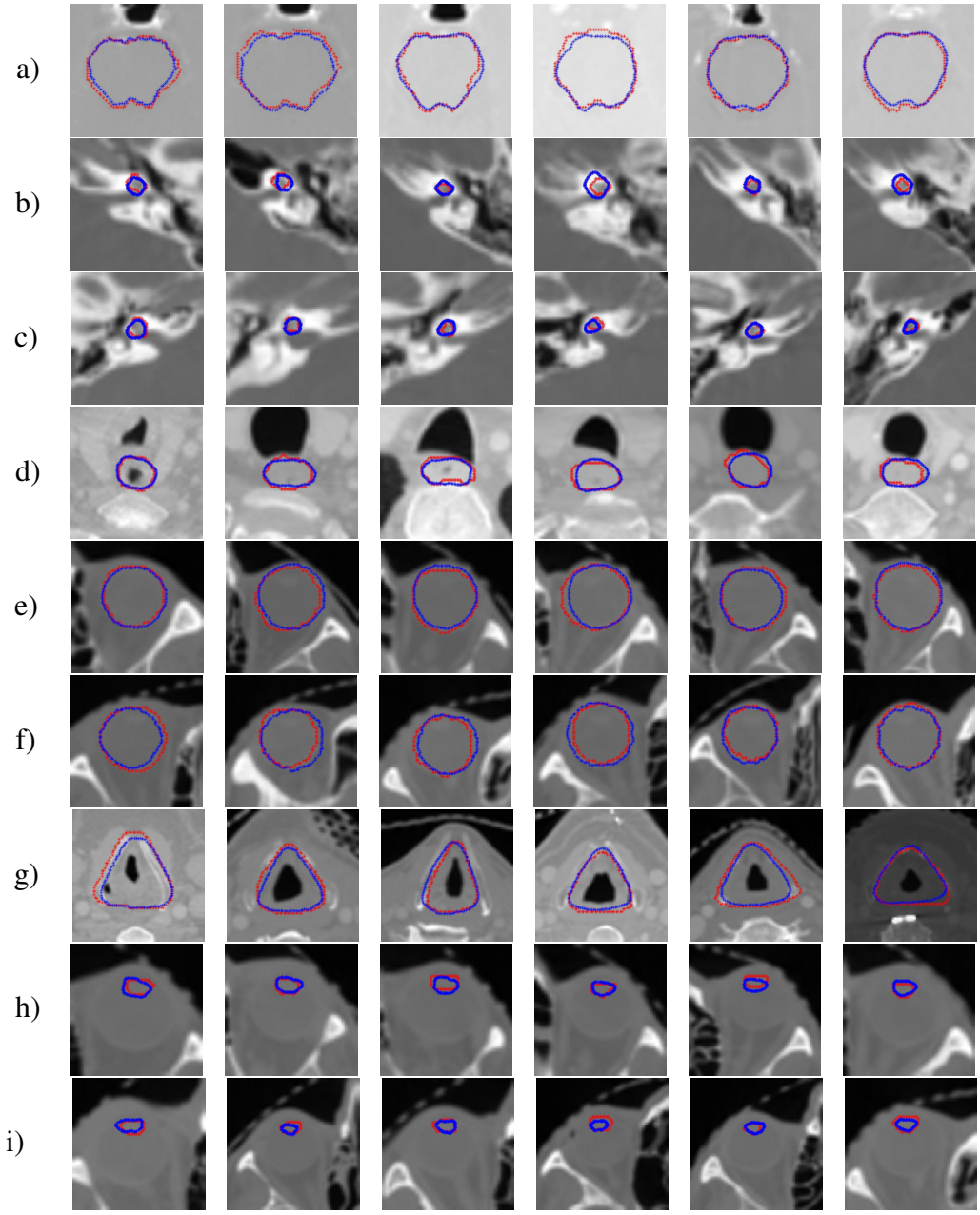


Figure 2.3. MSVR visual segmentation results part 1: a) brainstem, b) left cochlea, c) right cochlea, d) esophagus, e) left eye globe, f) right eye globe, g) larynx, h) left lens, and i) right lens. The approximate ground truth is shown with the red asterisks (*) and the automated segmentation results is delineated with the blue asterisks (*).

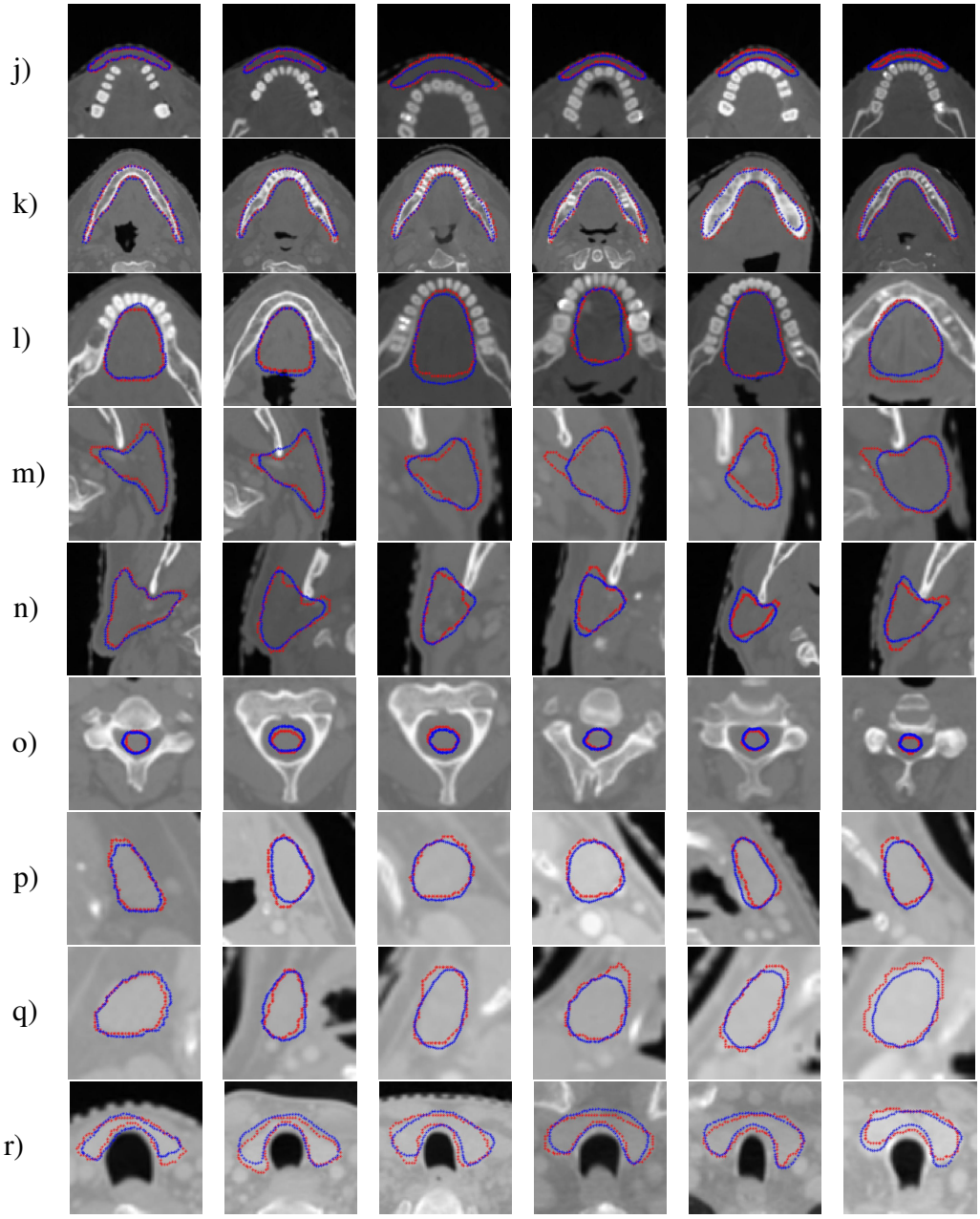


Figure 2.4. MSVR visual segmentation results part 2: j) lips, k) mandible, l) oral cavity, m) left parotid, n) right parotid, o) spinal cord, p) left submandibular, q) right submandibular, and r) thyroid gland. The approximate ground truth is shown with the red asterisks (*) and the automated segmentation results is delineated with the blue asterisks (*).

Our proposed model utilizes MSVR along with the HOG feature extractor to perform segmentation focusing on the anatomical structure or shape of the organs of interest. The model leveraged advanced machine learning in a holistic fashion formulating a segmentation task into a boundary regression problem. The locations of all boundary points were simultaneously regressed allowing for the global shape to guide the points that were previously learnt from the training data. Our model's performance in comparison to the approximate ground truth demonstrated high accuracy, achieving a DSC above 80%, for the following organs: brainstem, esophagus, both eye globes, larynx, mandible, oral cavity, both parotid glands, and spinal cord. The thyroid, both lens, lips, and right cochlea achieved DSCs within the 70% range, with the left cochlea achieving the lowest score of 66.9% (See Table 2.1). Organs with lower DSC scores suggest limitations in our model for smaller organs or organs with greater anatomical variances. While DSC scores of 70% are not optimal for clinical standards, indicating that much development is still needed, improvement on corresponding organs compared with that in current literature (see Section 2.4.2) demonstrates a step closer to achieving a machine learning tool that will be considerable for clinical implementations.

The MSE values reported in Table 2.1 indicated that our model's predicted values fit very well to the approximate ground truth data of each OAR, given the low MSE results of less than 1%. In addition, the variance of the predicted data compared to the approximate ground truth data was from 2.3% for both eye globes to 7% for the left parotid gland, which further demonstrated the high accuracy of our model, as low percentage values are represented as having less residual variance (See Table 2.1).

2.4.2 Literature comparative analysis

Baseline values of corresponding organs presented in this chapter were compared. It is important to note that this is a direct comparison of methods on different datasets comparing the same organs. In general, deep learning-based methods such as CNN require very large datasets in order to achieve optimal results. A machine learning model, such as the one presented in this paper, is advantageous compared to the state-of-the-art CNN [5] method as it allows for a smaller training dataset to obtain similar or even superior results with greater efficiency. CNN takes into account every permutation of parameters thus requiring a longer processing time. Our model had a training processing time of 49.6 minutes for all OAR on a single CPU compared to the CNN coupled with MRF training time that took a total of 6.5 hours [5] for all OAR on a graphics processing unit. Our MSVR model allowed for the

selection of specific parameters to process datasets for training, allowing the flexibility for optimization. Furthermore, another advantage of our model is the point based boundary representation that allows for more flexibility in capturing complex shapes since there is less assumption. This is also shown in the methods presented by Wang et al. [8] and He et al. [9] in their use of MSVR for spinal structures. DSC scores for our proposed algorithm demonstrated a smaller range in accuracy values among corresponding organs reported in the literature comparison (See Table 2.2). Our model consistently ranged within 82% to 93% versus the CNN model's DSC scores ranging from 69% to 89% for the same organs. This demonstrated that our model is competitively comparable to existing automated segmentation algorithms in literature.

Table 2.2. Segmentation results of corresponding organs-at-risk between the multi-output support vector regression (MSVR) model and convolutional neural network (CNN) coupled with Markov random fields (MRF) [5]. Dice similarity coefficient (DSC) is given in terms of mean (± standard deviation).

Organs	**MSVR DSC (%)**	**CNN + MRF DSC (%) [5]**
Eye globe left	93.8 ± 2.7	88.4 ± 2.7
Eye globe right	93.7 ± 2.9	87.7 ± 3.7
Larynx	89.1 ± 5.2	85.6 ± 4.2
Mandible	85.2 ± 5.3	89.5 ± 3.6
Parotid left	82.5 ± 7.0	76.6 ± 6.1
Parotid right	82.1 ± 8.5	77.9 ± 5.4
Spinal cord	83.1 ± 8.5	87.0 ± 3.2
Submandibular left	87.2 ± 5.8	69.3 ± 13.3
Submandibular right	87.1 ± 5.5	73.0 ± 9.2

References

[1] W. H. Organization, "Locally advanced squamous carcinoma of the head and neck," *Review of Cancer Medicines*, pp. 1–8, 2014.

[2] L. J. Peters, B. O'Sullivan, J. Giralt, T. J. Fitzgerald, A. Trotti, J. Bernier, *et al.*, "Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: Results from TROG 02.02," *Journal of Clinical Oncology*, vol. 28, no. 18, pp. 2996–3001, 2010.

[3] C. Chu, J. De Fauw, N. Tomasev, B. R. Paredes, C. Hughes, J. Ledsam, *et al.*, "Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans," *F1000Research*, vol. 5, p. 2104, 2016.

[4] V. Pekar, S. Allaire, A. Qazi, and D. Jaffray, "Head and neck auto-segmentation challenge: Segmentation of the parotid glands," *MICCAI 2010: A Grand Challenge for the Clinic*, pp. 273–280, 2010.

[5] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Medical Physics*, vol. 44, no. 2, pp. 547–557, 2017.

[6] V. Pekar, S. Allaire, J. Kim, and D. A. Jaffray, "Head and Neck Auto-segmentation Challenge," *Midas Journal*, pp. 349–445, 2009.

[7] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, B. M. Dawant, *et al.*, "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015," *Medical Physics*, vol. 44, no. 5, pp. 2020–2036, 2017.

[8] Z. Wang, X. Zhen, K. Tay, S. Osman, W. Romano, and S. Li, "Regression segmentation for M^3 spinal images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 8, pp. 1640–1648, 2015.

[9] X. He, A. Lum, M. Sharma, G. Brahm, A. Mercado, and S. Li, "Automated segmentation and area estimation of neural foramina with boundary regression model," *Pattern Recognition*, vol. 63, pp. 625–641, 2017.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 1, pp. 886–893, 2005.

[11] F. Pérez-Cruz, G. Camps-Valls, E. Soria-Olivas, J. J. Pérez-Ruixo, A. R. Figueiras-Vidal, and A. Artés-Rodríguez, "Multi-dimensional function approximation and regression estimation," in *Artificial Neural Networks — ICANN 2002: International Conference Madrid, Spain, August 28–30, 2002 Proceedings* (J. R. Dorronsoro, ed.), pp. 757–762, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002.

[12] M. Sánchez-Fernández, M. De-Prado-Cumplido, J. Arenas-Garcia, and F. Pérez-Cruz, "SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2298–2307, 2004.

[13] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of Database Systems. Springer, Boston, MA*, 2009.

[14] A. Elisseeff and M. Pontil, "Leave-one-out error and stability of learning algorithms with applications," *Advances in Learning Theory: Methods, Models and Applications, NATO Science Series III: Computer & Systems Sciences, Volume 190*, pp. 111–130, 2003.

CHAPTER 3

The contents of this chapter is an original research article: C.M. Tam, D. Zhang, T. Peters, and S. Li. Holistic multitask regression network for multiapplication shape regression segmentation (Submitted to the Medical Image Analysis journal for December 2019).

3 HOLISTIC MULTITASK REGRESSION NETWORK FOR MULTIAPPLICATION SHAPE REGRESSION SEGMENTATION

3.1 Introduction

Localization and delineation of each target organ is an integral part in medical image analysis for pathology diagnosis, treatment, and therapy planning. Clinical assessments require analyzing large volumes of data to identify multiple anatomic structures in multiple planes from multiple anatomic regions using a combination of imaging modalities, such as MRI or CT. Automated algorithms are developing rapidly due to their advantages of enhancing productivity, having greater consistency, and reproducibility compared to manual labelling and delineation [1]. However, conventional methods are limited to specific organ structures and cannot be generalized for other tasks without a great deal of modifications. Therefore, there is a greater demand for multiapplication frameworks; but, the development of a reliable and accurate multiapplication model is exceedingly difficult. Due to the diverse range of applications, automated localization and segmentation (even for a single application) are still hindered by the following challenges [2–4] (see Fig. 3.1):

1) *Diverse geometry.* Organ size and shape vary substantially; especially in the presence of pathologies. The complexity only increases with the addition of inter-subject variability and the diversity of applications, as shown in Fig. 3.1(a).

2) *Image appearance variability.* Images differ in contrast, brightness, resolution and possess inhomogeneous pixel intensities (e.g. bone and soft tissue in CT versus MRI (see Fig. 3.1(b)). Organs can also appear as one solid object or contain multiple components in distinct image slices. This is dependent on the plane of view and on the location the image slice was taken (see Fig. 3.1(c)). This becomes more complex when various structures throughout the body are considered.

3) *Discriminative feature embedding.* Due to the ambiguity of the boundary for soft tissue organs in CT images, organs affected by pathologies or impeded by artefacts,

generating discriminative feature embeddings is intractable, as shown in Fig. 3.1(e-g).

4) *Inherent imbalanced data of large and small organs.* Imbalanced organ data exists for datasets with multiple organs that span multiple anatomical regions. The quantity and distribution of images for a specific structure is inherently dependent on its geometry, position, and appearance frequency in the body, as shown in Fig. 3.1(d).

5) *Mis-localization and classification of similar shaped structures.* Organs with similar characteristics in close proximity are difficult to differentiate amongst each other (see Fig. 3.1(h)).

6) *Scarce and insufficient labelled data*, which can result in overfitting.

3.1.1 Related works

The novelty of our work is demonstrated by comparing existing works for medical image processing. These works can be divided into five categories: 1) localization, 2) segmentation, 3) localization and segmentation, 4) regression segmentation, and 5) multitask segmentation.

Localization-based methods: Existing works that primarily focus on the localization or detection of organ structures aim to locate and classify a target organ, thus generating regions of interest used for further image processing and analysis. Methods in this category use machine learning, deep learning classifiers, or a combination of methods. An example of a localization-based work is described in [5] where the authors use a random forest classifier combined with volume descriptors to perform vertebrae localization in CT scans. In [6] support vector machine (SVM) based combined with Markov Random Fields (MRF) were used for vertebrae and intervertebral disc localization. The work in [7] used deep Convolutional Neural Networks (CNN) to detect vertebrae in MRI and CT images. In [8] SVM coupled with the Histogram of Oriented Gradients [9] were used for disc localization. The work in [10] performed vertebrae localization by incorporating local and global symmetry feature using a sphere surface expansion method and an iterative optimization framework. In general, localization or detection methods primarily serve as a pre-processing step to segmentation, which is one of the main focused aspects in this chapter.

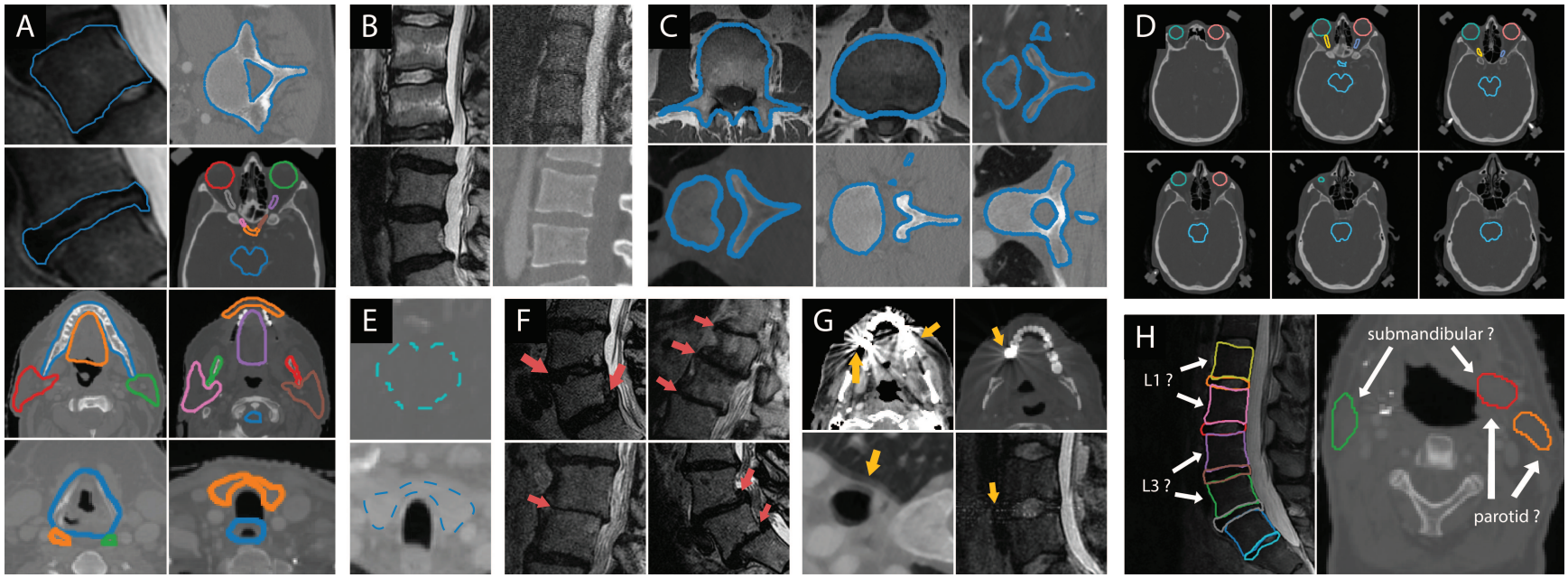


Figure 3.1. The challenges of multiapplication segmentation. (a) Diverse geometry; (b) appearance variability in resolution and contrast; (c) appearance variability among slices in the same plane of view; (d) inherent imbalance of samples due to structure characteristics; (e-g) ambiguous boundary due to low contrast of soft tissues in CT images, pathologies (indicated by the red arrows), and artefacts (indicated by the yellow arrows); (h) mis-localization and classification of similar shaped structures.

Segmentation-based methods: Some methods directly focus on the segmentation task skipping the detection or localization process all together. For instance, some of the methods currently used in literature for segmentation consist of atlas-based methods such as in [4, 11, 12] for head and neck organs in CT images, atlas-based coupled with machine learning [3, 13], model-based [14] for head and neck CT image segmentation, in [15] a graph cut framework was used to segment the optical nerves, and in [16] graph cuts was used to segment the intervertebral disc. More recently, deep learning methods are being more widely used and improved for semantic segmentation in medical image analysis, such as the current state-of-the-art U-Net developed by [17]. Semantic segmentation methods perform pixel-wise classification that gives semantic context to each pixel in the entire image. [18] proposed two architectures based on U-Net for segmentation, a recurrent convolutional neural network (RU-Net) and a recurrent residual convolutional neural network (R2U-Net). The two models were validated on three datasets (blood vessels in retina images, skin cancer, and lung lesion). [19] trained a U-Net with dilated convolutions to perform semantic segmentation for head and neck tumours.

Localization and Segmentation-based methods: These methods aim to combine both the localization and segmentation task in a single framework, similar to our proposed method. For example, in [2] a four-step deep learning method was used combining a CNN and MRF to classify and segment organs in head and neck CT images (brainstem, optic nerves, chiasm, parotid glands, submandibular glands, eye globes, pharynx, larynx, mandible, and spinal cord). The four-step method included voxel classification using CNN, smoothing with MRF, removal of small components, and dilate-erode operations to achieve the final segmentation results. In [20] a model-based method was used for the detection, classification, and segmentation of vertebral bodies in CT images. In [21] an Adaboost joint method with an iterative normalized cut algorithm was used for vertebrae detection and in [22] a Gabor filter bank-based method was used for intervertebral disc localization and segmentation in MRI images. [23] adapted the U-Net architecture to perform two-stage 3D multiclass cardiac segmentation.

Shape Regression-based methods: Shape regression methods, also known as boundary regression, have shown comparable or exceptional performance compared to conventional methods for medical image segmentation. Shape regression methods utilizes a target object's global shape to predict the points along the new target's contour. Furthermore, since the boundary representation for the following works are point-based, shape regression methods possess the flexibility to adapt to several diverse types of geometry. In [24] a

multi-kernel multi-output support vector regressor was used to regress the boundary points of multi-plane, multi-modality, and multi-structure (M^3) spinal images. In [25] CNN was used for the boundary regression of cardiac structures following the same pre-processing protocol for the data as in [24].

Multitask-based methods: Multitask learning algorithms incorporate the benefits from several related tasks to improve the overall performance by taking the underlying common information that may be ignored by single task learning algorithms. [26] designed a multitask shape regression (MTSR) model to simultaneously estimate coordinate points on shape contours and model coordinate correlations. The MTSR architecture used HOG for feature representation and modelled inherent correlations via sparse learning. [26] further demonstrated the capabilities of shape regression by extending its application to six different representative datasets (knee, clavicle, prostate, cardiac bi-ventricles, cardiac 4 chambers MR, and cardiac 4 chambers CT). [27] introduced a U-Net-like encoder-decoder architecture to simultaneously segment thoracic organs and perform global slice classification on CT scans. [28] developed a multitask U-Net to solve three tasks: object instance detection, separation of wrongly connected objects, and semantic segmentation for food microscopy images.

While all these aforementioned methods are effective for each of their respective applications, they all have at least one of the following limitations: 1) one specific imaging modality, 2) one specific plane, 3) one specific anatomical structure, and 4) one specific application. Task-specific algorithms have the advantage of leveraging the inherent characteristics of a target structure's anatomy and unique surroundings, such as the number of visible objects, their individual instances in a given image, and their anatomical identity. However, a general or multitask algorithm cannot use such specific assumptions or prior knowledge. It considers individual object instances as independent from one another; similar to the instance segmentation methods for natural images. Conversely, similar to single task methods multitask segmentation methods are also designed with respect to their tackled applications, but possess increased generalizability. If a multitask algorithm were, however, trained on more variable applications the multitask model would then be capable of generalizing to even more diverse tasks.

3.1.2 Overview of the proposed method

In this study, a holistic multitask regression network (HMR-Net) for multiapplication image analysis is proposed to tackle the segmentation and localization challenges (see Fig.

3.2). HMR-Net formulates segmentation into a multitask regression problem, which leverages the strength of jointly solving multiple tasks. HMR-Net performs object instance detection and shape regression with its flexible boundary representation enabling the ability and potential to tackle multi-object and multiple applications as seen in [26]. Manifold embedding was incorporated to encode holistic shape information by modelling coordinate correlations. Simultaneously, HMR-Net unravels the nonlinear relationships between image appearance and distinct object properties with hierarchical multiscale and fused features.

HMR-Net's framework seamlessly combines four sub-modules: 1) Two N-residual networks (N-ResNets) with cross-stitch units was implemented for multiscale and fused feature representations. 2) A region proposal network (RPN) to highlight regions of interest (ROI). 3) A linear regression network (LRN) for multiclass object detection. 4) A shape regression network (SRN) to directly estimate the boundary coordinates of each ROI.

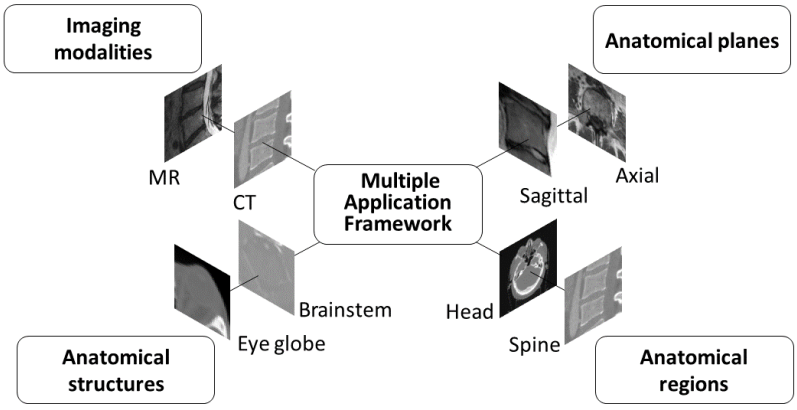


Figure 3.2. The proposed holistic multitask regression network (HMR-Net) tackles the diversity of medical images in multiple applications with a single framework.

- To tackle the *diverse geometry challenge*, an organ's shape was represented as individual boundary coordinates predicted using the SRN; a simple convolutional regression network. The SRN performs simultaneous and direct regression of each corresponding boundary point. As individual coordinates, the points successfully capture a variety of geometric shapes; making it an ideal choice to tackle geometric variability.

- To solve the *image appearance and discriminative feature embedding challenge*, multiscale and fused features were generated to represent medical images using N-ResNet. Coordinate information about the data was directly integrated into the fea-

ture extraction pipeline to preserve and enhance precise spatial information. Furthermore, coordinate correlations were modelled to handle boundary ambiguity. For organs appearing as a single object or with multiple pieces, the coordinate boundary representation is capable of handling multi-component organs.

- To handle the *inherent imbalanced data of large and small organs challenge*, augmented resampling was used to balance the majority and rare class samples. Data augmentation would only be applied to samples containing the rare and minor classes.

- To resolve the *mis-localization of similar shaped structures challenge*, multitask learning of features from the RPN, LRN, and SRN act as a form of attention focusing mechanism enabling the network to discard irrelevant features and better retain relevant features for all tasks.

- To handle the *scarce and insufficient labelled data challenge*, the use of data augmentation with slight translations, rotations, random cropping and flipping, Gaussian blurring, and scaling helped to reduce the potential of overfitting.

3.1.3 Contributions

The main contributions are as follows:

1) **Application:** Achieved multiapplication medical image shape regression segmentation in a single framework.

2) **Approach:** Formulated segmentation into a multitask regression problem to leverage deep learning in a holistic fashion.

3) **Methodology:** Proposed a novel multitask regression deep learning framework with multiscale + fused feature representation enhanced by task correlations.

The rest of this chapter is organized as follows: Section 3.2 describes the network's architecture and each of the separate sub-modules. Section 3.3 presents the experiment details, such as the validation datasets and evaluation metrics. Lastly, results and detailed discussions are in Section 3.4.

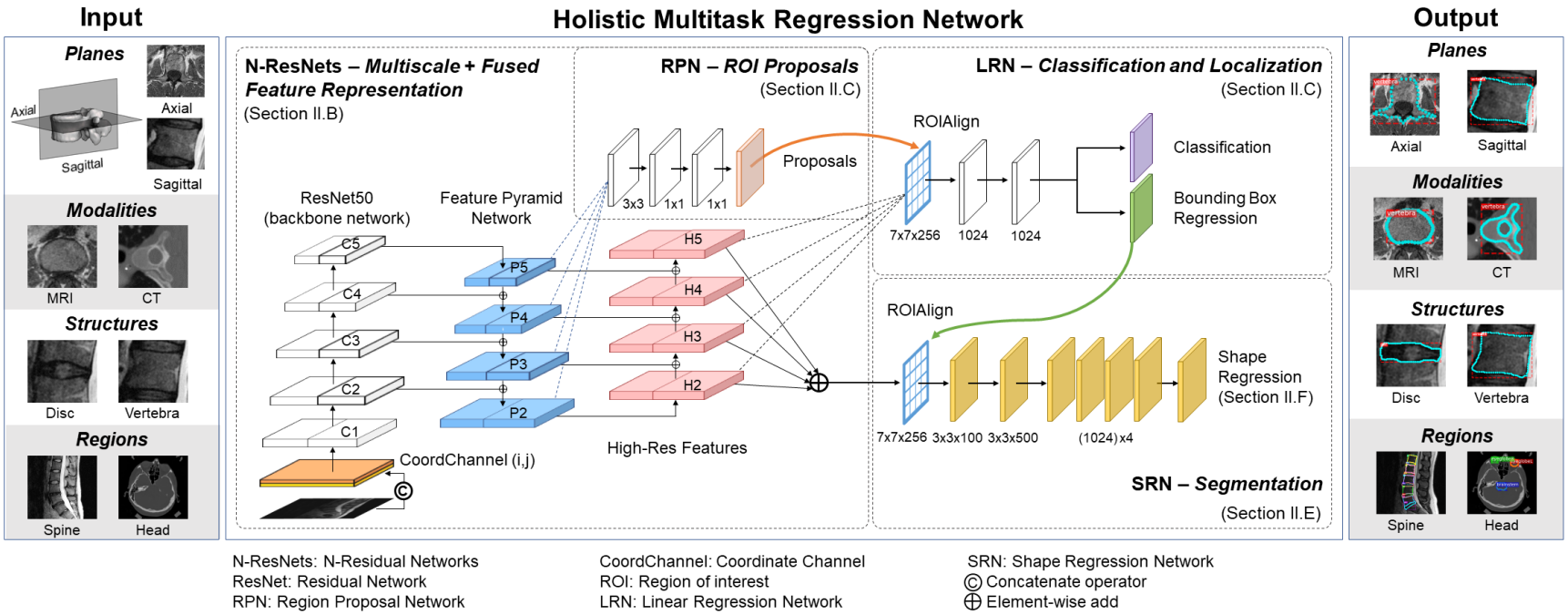


Figure 3.3. HMR-Net's architecture. Two N-residual networks (N-ResNets) with cross-stitch units for multiscale and fused feature representation, a region proposal network (RPN) for coarse region of interest (ROI) prediction, a linear regression network (LRN) for multiclass object detection, and a shape regression network (SRN) for regression segmentation. Layers with a line through them (*C1-H5*) represent layers joined with cross-stitch units.

3.2 Holistic Multitask Regression Network (HMR-Net)

3.2.1 Multitask regression

The proposed HMR-Net (shown in Fig. 3.3) formulates the segmentation task as a holistic multitask regression problem to leverage joint feature learning between tasks. HMR-Net simultaneously learns three tasks (classification, localization, and shape regression segmentation), which enables the network to learn task correlated features preferable to each task. The joint learning of multiple tasks allows the model to focus on more relevant features supported by the other tasks and enhances the model's ability to generalize to new tasks in similar environments. HMR-Net's feature parameters in N-ResNet are shared among the three task-specific sub-modules (RPN, LRN, and SRN).

Multitask regression is also applied during shape regression, where the estimation of each point is a regression task. HMR-Net explicitly models coordinate correlations in a similarity matrix incorporated into the target output space (see Section 3.2.5). Holistic shape information is captured by modelling the inherent correlations among points.

3.2.2 Multiscale and fused feature representation by N-ResNet with cross-stitch units

The representation network, designed as two N-residual networks (N-ResNets) with cross-stitch units, combines local and global features to handle the nonlinear relationship between image appearance and object properties. A single N-ResNet incorporates a residual network (ResNet) and feature pyramid network (FPN) (adopted from [29]) as part of its backbone for multiscale feature representations, but additionally combines a new bottom-up pathway with fused features merged together from all levels of the network. ResNet is one of the widely used networks for regression type problems and is proven to be compatible with the FPN architecture in both [29, 30].

The FPN extracts ROI features from distinct levels of the feature pyramid according to their scale. As output, different scaled feature maps are generated to provide multiscale organ feature representations. The formula from [31] to perform feature map selection based on the width w and height h of an ROI is given below:

$$k = [k_o + log_2(\sqrt{wh}/224)], \tag{3.1}$$

where k is the Pk layer in the FPN and k_o is the target pyramid level on which to map an

ROI with $w \times h = 224^2$. Here 224 is ImageNet's pre-training size and k_o is set to 4 as in [31]. Intuitively, Eq. 3.1 means that if an RoI's scale is smaller, for example 1/3 of 224, it should be mapped into a finer-resolution level, such as $k = 3$.

Where the N-ResNet backbone differs from that of the ResNet+FPN is that a new bottom-up pathway with clean lateral connections is introduced. This bottom-up pathway was motivated by the necessity to further enhance localization of the entire feature hierarchy for classification and boundary regression. The constructed pathway was inspired by [32], in that high responses to edges is required to accurately locate object instances. Therefore, a bottom-up path to propagate strong responses of low-level patterns to top-level layers was constructed for more precise classification and regression. HMR-Net further enhances the retention of local features by generating higher resolution feature map outputs for each layer of the FPN for classification.

Bottom-up pathway structure: The bottom-up pathway begins from the lowest FPN level P2 and ends at P5, as shown in Fig. 3.3. From P2 to P5 the layers are gradually up-sampled to the size of $P2$ (128×128×256). The new layers are denoted as $H2, H3, H4, H5$. For each building block a higher resolution feature map H_i and a coarser map P_{i+1} are added by lateral connections and processed by a 3×3 convolutional layer to generate the new feature map H_{i+1}. A filter size of 256 was consistently used in these building blocks. The feature map for each proposal are then pooled from these new feature maps. Separately for the boundary regression sub-module, the higher resolution features are concatenated into a single layer before being pooled by the ROIAlign layer (see Fig. 3.3).

In addition, a coordinate convolution layer (CoordConv from [33]) was implemented to directly incorporated coordinate (i, j) data into the input channels. This allows the convolutional filters to discern their location in the Cartesian space for further spatial precision. As indicated in Fig. 3.3 by CoordChannel, pixel coordinates were added as two additional dimensions as part of the input data.

Cross-stitch units: Cross-stitch units enable the multitask network to automatically learn the optimal combination of task-specific and shared representations. Inspired by [34] and [35] cross-stitch units were used to combine two N-RestNet architectures. Similar to in [34] two networks are joined together; however, the alpha parameters are not manually set but are learned as describe in [35] for their sluice networks. In the proposed HMR-Net, the alpha parameters allow the network to learn whether to share or focus on task-specific

features in a subspace.

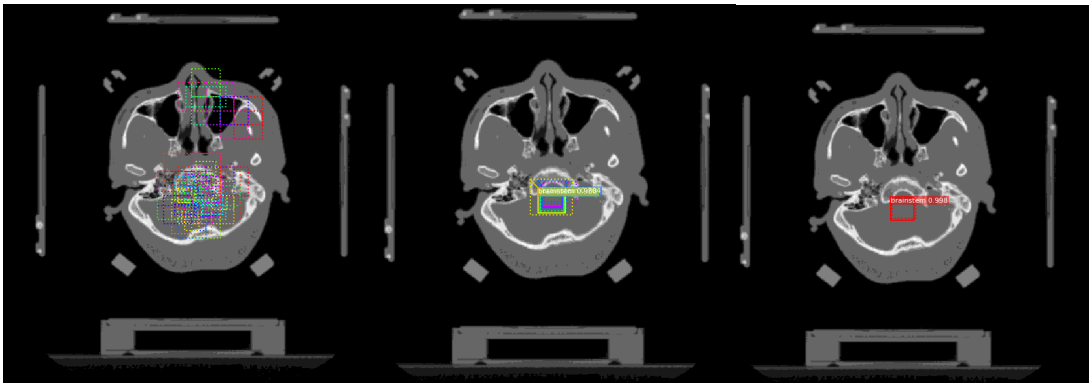
While the concept of multitask networks may have many benefits as discussed in Section 3.2.1, some tasks when joined together degrades the overall performance. This is due to the respective tasks not being closely related. This phenomenon is called negative transfer. Loosely related tasks may only benefit from shared parameters at the beginning but suffer when sharing information in the later layers. The amount to share would vary per task, which makes pre-set alpha values infeasible.

Cross-stitch units provides a solution to negative transfer, thus allowing the network to learn what to share between loosely related tasks. Shape regression and detection are only loosely related tasks. Shape regression requires a much finer precision to directly estimate individual coordinates along a organ's shape contour. The exact values and behaviour of the alpha parameters were not monitored in this current study as it was outside the scope of this paper. Studying the fluctuations of the alpha values would provide a lot of insight into the learning behaviour of deep neural networks, which will be explored in a future study.

3.2.3 Coarse-to-fine organ localization by RPN and LRN

Organ classification and localization was performed using a two-stage coarse-to-fine scheme. A region proposal network (RPN) and a simple linear regression network (LRN) was re-implemented as described in [29] and [30] for bounding box regression and classification regression. However, a maxout [36] activation was used for the final fully connected layer. Following [29], a ROI alignment (ROIAlign) layer instead of the ROI pooling layer (referred to as ROIPool in [30]) was used for greater spatial accuracy. The ROIAlign layer aligns the extracted features with the input image using bilinear interpolation. ROIAlign computes the exact values of the input features with four nearby sampled locations in each ROI, thus preserving pixel-to-pixel spatial correspondence.

In the coarse-to-fine scheme, RPN performs coarse ROI localization with three stages (see Fig. 3.4): 1) ROI estimation with a series of candidate bounding boxes, 2) bounding box refinement to remove boxes with no ROI, 3) non-maximum suppression to remove low scoring duplicate boxes. Then LRN performs fine ROI localization using the outputs from RPN. LRN extracts features using ROIAlign from each candidate box and performs classification and bounding-box regression.



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

Figure 3.4. Visualization of the three stages in the region proposal network: (a) regional proposal to propose candidate boxes, (b) bounding-box refinement to remove boxes with no object, and (c) non-maximum suppression to remove low scoring duplicate boxes.

3.2.4 Boundary representation

For each organ boundary, the coordinate points were generated from approximate ground truth segmentation masks, where p points were evenly sampled along an organ's boundary. This form of boundary representation provided the necessary flexibility to address organ shape, size, and image appearance variability in medical images. Each organ's boundary was represented by a set of p points, $boundary_{(i)} = (x_i, y_i)|_{i=1:p}$. (x_i, y_i) is the i^{th} boundary point's coordinates. The value of p determines the resolution of the boundary shape. Too small a p value will result in a loss of detail in an organ's local boundary. However, too large a p will increase the complexity in the network. We determined from literature [24, 25, 37, 38] and from experimenting with different values that $p = 50$ for full-scaled images (resolution of 512x512 or more) and $p = 100$ for cropped/patch images smoothly approximated complex organ shapes without increasing complexity. A scaled normalization step was performed to all input images to ensure the points are in a consistent coordinate system across different images. Additionally, a rescaling step was required to transform the predicted boundary points back to the original image space during testing.

3.2.5 Shape regression segmentation by SRN

Shape regression directly estimates the coordinate points along the shape contour of an organ [26]. HMR-Net performs shape regression by multitask regression. HMR-Net simultaneously predicts the coordinates while jointly capturing point correlations to attain holistic shape information. The captured shape information can be utilized by HMR-Net

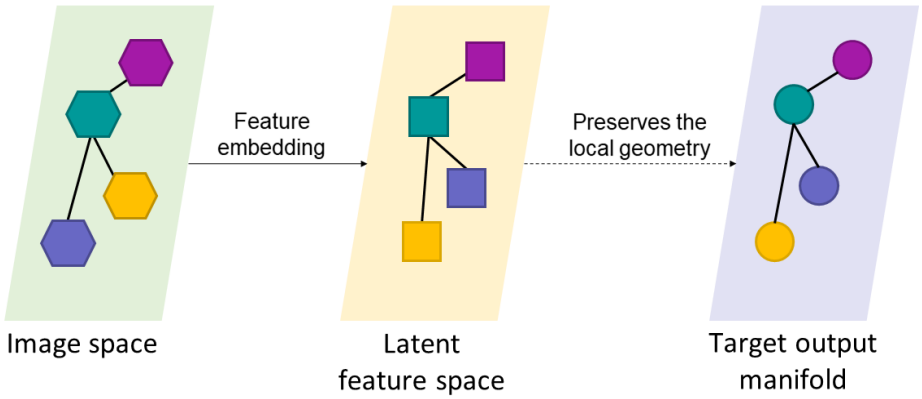


Figure 3.5. Manifold learning. HMR-Net preserves the intrinsic local geometry of the data by mapping a latent feature space into the target output manifold. The points lying in the three spaces with the same colour are the same sample.

as a template to predict new instances in new samples, or even to recover contours in the absence of clear edges and region homogeneity. A segmentation mask can then be generated from the predicted contour.

The SRN, as shown in Fig. 3.3, consists of two convolution layers with filter sizes of 100 and 500 respectively, and four fully connected layers with filter sizes of 1024 each. Batch normalization is used after each convolutional layer followed by a ReLu activation layer. The maxout [36] activation was used for the final fully connected layer.

Manifold regularization: Inherent spatial and statistical correlations among each boundary point was modelled by implementing manifold regularization to learn the shared features among correlated points. A latent feature space (latent manifold) mapping the coordinate correlations of the input image was embedded into the target space (target output manifold), as shown in Fig. 3.5. The motivation for manifold regularization is based on the idea that neighbouring samples on the target output manifold are similar in the latent manifold [39]. Manifold regularization preserves the intrinsic local geometry of the data, which is used to regularize and supervise the training of HMR-Net's shape regression.

The latent feature space (see Fig. 3.5) is represented by a constructed k-nearest neighbour graph $G = (V, E)$, where V and E are the vertices and the edges between the vertices, respectively. The graph is constructed on multivariate targets $\{Y_{(i)}\}_{i=1}^{N}$, where $Y_{(i)} = \{y_1, \ldots, y_P\}|_{P=2p}$ and N is the number of image samples, to represent the local neighbour relationship. A similarity matrix $S \epsilon \mathbb{R}^{N \times N}$ (also known as an adjacency matrix) was implemented to explicitly model coordinate correlations for each batch of images. The similarity

matrix is a symmetric $N \times N$ dimensional matrix consisting of non-negative elements that correspond to the edge weights of the graph G. Each element $S_{i,j}$ was calculated by a heat kernel with a parameter $\sigma = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1,j>1}^{N} \|Y_i - Y_j\|_2$:

$$S_{i,j} = \begin{cases} \exp\left(-\frac{\|Y_i - Y_j\|_2^2}{2\sigma^2}\right), Y_j \epsilon N_k(Y_i), i \neq j, \{i,j\} = 1,2,\ldots,N \\ 0, otherwise \end{cases} \tag{3.2}$$

where $\|\cdot\|_2$ is the l_2-norm and N_k obtained from the Euclidean norm represents the k-nearest neighbours of Y_i. The l_2-norm was adopted as it provides a good approximation of the distance of the shortest path between two points for the neighbouring points in a manifold [39].

The graph G, constructed using Eq. 3.2, is an asymmetric directed graph, which was replaces with a symmetric similarity matrix $S \epsilon \mathbb{R}^{N \times N}$ to create an undirected graph G for easier computation: $S = max(S, S^T)$, where $max(\cdot)$ denotes the element-wise maximum operation and T is the transpose operation. The manifold regularization loss is defined as follows:

$$L_{manifold} = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{N} u(S_{i,j})} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{i,j} \|z(x_i) - z(x_j)\|_2^2 \tag{3.3}$$

$$u(t) = \begin{cases} 1, t > 0 \\ 0, t \leq 0 \end{cases} \tag{3.4}$$

The manifold loss in Eq. 3.3 was inspired by the work in [39], which used the principles of Laplacian Eigenmaps [40]. A penalty would incur when similar vertexes are mapped far away in the latent space [39]. For our purposes, constructing a Laplacian graph would aid in minimizing sparse or outlier points along an organ's shape contour. A Laplacian matrix was constructed using the l_2-norm, which is optimizable using stochastic gradient descent [41]. Contrast to the manifold loss in [13] the normalized Laplacian was used where,

$$normL = D^{-1/2} L D^{-1/2}. \tag{3.5}$$

$D \epsilon \mathbb{R}^{N \times N}$ is the diagonal matrix. L represents the unnormalized Laplacian given by $L = D - S$, and S is the similarity matrix. For the normalized Laplacian, the eigenvectors are in

a "normalized" form, which are more consistent with the eigenvalues in spectral geometry and in stochastic processes. The choice to implement the normalized Laplacian compared the unnormalized Laplacian was based on the notion by [42] that the normalized Laplacian can be generalized to all graphs. The normalized Laplacian gives each vertices a weight proportional to its degree, which can lead to better results [43].

The manifold regularization loss with the constructed normalized Laplacian graph is rewritten as follows:

$$L_{manifold} = \frac{2}{\sum_{i=1}^{N}\sum_{j=1}^{N} u(S_{i,j})} tr(Z^T normLZ), \tag{3.6}$$

where $Z \epsilon \mathbb{R}^{N \times M}$ are the samples in the latent space and $tr(\cdot)$ is the trace operation of a matrix.

3.2.6 Multitask loss function

The optimization function for the HMR-Net is a hybrid multitask loss function, which enables the network to learn its three tasks (classification, localization, and segmentation) by regulating the weights and biases that will benefit each task. The first two components of the loss function used in this paper is the same as the multitask loss function as implemented in [29]. However, the mask loss is replaced with a shape regression loss. Our holistic multitask loss function for an image is defined as:

$$\begin{aligned} L_{multitask}(w_i) = \lambda_c \sum_i L_{class}(w_i) + \lambda_b \sum_i L_{bbox}(w_i) + \\ \lambda_s \sum_i L_{shape}(w_i) + \lambda_m \sum_i L_{manifold}(w_i) + \lambda_{reg} \sum_i ||(w_i)||_2, \end{aligned} \tag{3.7}$$

which is the sum of the loss function for the classes L_{class}, the loss function for the bounding boxes L_{bbox}, the loss function for the boundaries L_{shape}, and the manifold regularization loss $L_{manifold}$ (see Eq. 3.6) as shown above. The manifold loss helps to regularize the estimated points to remain inside their real distribution. $\lambda_c, \lambda_b, \lambda_s$, and λ_m, are the scaling factors for their corresponding importance in the loss. The last term with is the l_2-norm regularization for the trainable weight w_i and λ_{reg} is a hyperparameter. Note L_{bbox} and L_{shape}, the bounding box loss and shape regression loss respectively, are activated only for positive anchors as described in [30].

The class loss function employed was the softmax cross-entropy for multi-label classification. Whereby p_i is the predicted probability of anchor i being an object and p_i^* is the approximate ground truth label (binary).

$$L_{class}(p_i, p_i^*) = -p_i^* log(p_i) - (1 - p_i^*) log(1 - p_i), \text{ or}$$

$$L_{class}(p_i, p_i^*) = \begin{cases} -log(p_i) & \text{if } p_i^* = 1 \\ -log(1 - p_i) & \text{otherwise} \end{cases} \tag{3.8}$$

Smooth-L1 loss was used for the bounding box loss function and multi-organ localization. t_i is the vector of 4 parameterized coordinates of predicted bounding boxes and t_i^* is the approximate ground truth bounding box associated with a positive anchor.

$$L_{bbox}(t_i, t_i^*) = \sum_{i \epsilon \{x,y,h,w\}} \text{smooth}_{L1}(t_i, t_i^*)$$

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3.9}$$

The mean squared error (MSE) combined with the Euclidean distance metric was selected for the shape regression loss function. q_i is the vector of N parameterized coordinates of predicted boundary points and q_i^* is the approximate ground truth boundary associated with a positive anchor.

$$L_{shape}(q_i, q_i^*) = \frac{1}{N} \sum_{i=1}^{N} \left((q_i^* - q_i)^2 + \|q_i\|_2^2 \right), \tag{3.10}$$

Stochastic gradient decent was used as our optimization function and a L2 regularization was added to further avoid overfitting the model.

3.3 Experiment Configurations

3.3.1 Datasets

In this chapter, three regions of the body that contains the most organ variability were examined: head, neck, and spine. These regions validate our model's capabilities of adapting

Table 3.1. Statistics of the eight datasets.

Dataset	Structures	Subjects	Images	Planes	Modalities
Spine (SD-MR-Sag)	Intervertebral Disc	93	465	Sagittal	MR
Spine (SV-MR-Sag)	Vertebral body	93	465	Sagittal	MR
Spine (S-MR-Sag)	Intervertebral Disc / Vertebral bodies (L1, L2, L3, L4, L5)	93	93	Sagittal	MR
Spine (SV-MR-Ax)	Vertebral body	10	50	Axial	MR
Spine (SV-CT-Sag)	Vertebral bodies	10	170	Sagittal	CT
Spine (SV-CT-Ax)	Vertebrae	10	170	Axial	CT
Head	Brainstem / Chiasm / Eye globe (LR) / Optic nerve (LR)	53	5164	Axial	CT
Head and Neck (HaN)	Brainstem / Eye globe (LR) / Esophagus / Larynx / Lips / Mandible / Oral cavity / Parotid (LR) / Submandibular (LR) / Thyroid	56	10776	Axial	CT

Notation: (LR) refers to the left and right structure belonging to the specified organ class.

for different applications. The proposed model was trained and evaluated on eight separate task datasets generated from five main image sets. The statistics of the task datasets are reported in Table 3.1. The first image set consisted of head CT images obtained from 53 anonymous patients. The second image set consisted of head and neck CT images obtained from 56 anonymous patients, and lastly, three spine datasets of MRI and CT images were collected from 113 anonymous patients (including disc degeneration and vertebrae fracture cases). The total number of patients from all five image sets was 222. All CT images in the first two image sets were axially reconstructed. Images from the head dataset came from different CT scans with varying pixel spacing settings for each patient. Images from the head and neck dataset had a pixel width and height of 1.27 mm, voxel depth of 1.25 mm, and a matrix dimension of 512×512. The CT spine datasets contained regions of interests cropped manually. The first MRI spine image set consisted of sagittal T2-weighted 1.5 T MRI scans from 93 subjects (46 men, 47 women, average 49 ± 15 yrs) with a TR (repetition time) of 4000 ms and TE (echo time) of 85 ms. The in-plane resolution was 0.5 mm $\times$ 0.5 mm with a slice thickness of either 1 mm or 1.6 mm. The scans covered the entire lumbar spine and a small portion of the thoracic spine. The second MRI spine image set consisted of axial T1-weighted 1.5 T MRI scans from 10 subjects (6 men, 4 women, average 48 ± 9 yrs) with a TR of 500 ms and TE of 11 ms. The in-plane resolution was 0.4 mm $\times$ 0.4 mm with a slice thickness of 4 mm. The scans covered the entire lumbar spine. The CT spine image set is publicly available at http://csi-workshop.weebly.com/challenges.html [44] consisting of CT scans from 10 young adults (16-35 yrs old). The in-plane resolution was between 0.31 and 0.45 mm and the slice thickness was 1 mm. The scans covered the entire thoracic and lumbar spine.

Manual approximate ground truth delineations were performed by experienced radiologists for the following organs: brainstem, chiasm, cochleae, esophagus, eye globes, larynx,

lenses, lips, mandible, optical nerves, oral cavity, parotid glands, spinal cord, vertebrae and intervertebral discs from the lumbar and thoracic regions, submandibular glands, and thyroid. The average of the manual annotations between two physicians were used as the approximate ground truth. Since deep learning methods requires large amounts of data, augmentation was performed on the training samples to improve the generalization performance by randomly flipping, rotating, translating, scaling, cropping, and adding a Gaussian blur.

3.3.2 Training configurations

Table 3.2. HMR-Net's training run time.

Datasets	Run Time (hours)
SD-MR-Sag	9.9
SV-MR-Sag	10.7
SV-MR-Ax	12.9
SV-CT-Sag	8.2
SV-CT-Ax	11.8
S-MR-Sag	27.5
Head	56.4
HaN	61.2

The proposed model was implemented in Python 3.6, Keras 2.1.5 [45], and a TensorFlow 1.9 library [46]. Training was conducted on Nvidia P100 Pascal and Nvidia GTX 1080ti GPUs. The initial learning rate was 0.001 with a decay rate of 0.000001 and batch size of 3. Weight decay and Nesterov momentum were 0.0001 and 0.9 respectively. The aforementioned values are the initial hyperparameter settings to train the eight datasets. The hyperparameters for smaller size datasets or imbalanced datasets were manually adjusted to better regularize overfitting, by adding spatial dropout values and kernel regularization. *k*-fold cross-validation (detailed in Section 3.3.3) was used to optimize the training hyperparameters. The ResNet+FPN backbone and RPN re-implementations were based on TensorFlow and Keras using [47]. The training run times are displayed in Table 3.2 for each of the eight task datasets on single GPUs. We summarize HMR-Net's training algorithm in Table 3.3.

Table 3.3. Summary of HMR-Net's training algorithm

Algorithm 1: HMR-Net Training

Input: an image $I = \{I_1, ..., I_N\}$ with N pixels, class labels,
bounding boxes, boundary coordinates, and a similarity matrix (Section 3.2.5):
$Y_{class} = \{c_1, ..., c_N\}$, $Y_{bbox} = \{y_1, x_1, y_2, x_2\}$,
$Y_{boundary} = \{y_1, x_1, ..., y_p, x_p\}|_{p=50 \ or \ p=100}$,
$Y_{manifold} = S \epsilon \mathbb{R}^{N \times N}$

Output: $\mathbf{w}_{nresnet}, \mathbf{w}_{rpn}, \mathbf{w}_{lrn}, \mathbf{w}_{srn}, \mathbf{w}_{rpnclass}, \mathbf{w}_{rpnbbox}, \mathbf{w}_{class}$,
$\mathbf{w}_{bbox}, \mathbf{w}_{boundary}$

1 He normal initialization of weights;
2 **while** $L_{multitask}$ not converged **do**
3 normalize Y_{bbox} and $Y_{boundary}$ coordinates;
4 compute (3.1) for ROI feature map selection;
5 compute $L_{multitask}$ (3.7);
6 update $\mathbf{w}_{nresnet}$, $\mathbf{w}_{rpn}$, $\mathbf{w}_{lrn}$, and $\mathbf{w}_{srn}$ with back propagation from $L_{multitask}$;
7 **end**

3.3.3 Evaluation metrics

Five evaluation metrics were used in this study to evaluate the performance of our model: 1) k-fold cross-validation, 2) the intersection over union (IoU), 3) mean average precision (mAP), 4) dice similarity coefficient (DSC), and 5) the Hausdorff distance (HD).

A k-fold cross-validation scheme was used to split our dataset for training and testing. The cross-validation computes a statistic on the left out samples. For this metric k is the number of equal sized subsamples that the original sample is randomly partitioned into; in this paper $k = 5$. For each k iteration (folds), the model is trained with $k - 1$ subsamples leaving a single subsample to be used for testing [48].

IoU (also known as the Jaccard index) was used to measure the amount of overlap between the approximate ground truth bounding boxes and the automatically generated bounding boxes. IoU is given by the following expression,

$$IoU(M, A) = \frac{M \bigcap A}{M \bigcup A}, \tag{3.11}$$

where M is the area of the approximate ground truth bounding box, A is the area of the automated bounding box, $M \bigcap A$ is the area of the overlap region between M and A, and $M \bigcup A$ is the area of union of the approximate ground truth and automated bounding boxes combined. IoU has a range of (0,1) where the change from 0 to 1 represents no overlap to complete overlap. The threshold for an IoU > 0.7 indicated a positive match and IoU < 0.3 for negative matches.

Localization accuracy was evaluated using mAP, which is the average of all the average precisions from all the classes in the dataset over IoU thresholds of 0.5 to 0.95. mAP was taken over a range of confidence thresholds to provide an overall view of the whole precision and recall curve for the model. Precision and recall are given by: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, respectively, where TP is true positive, FP is false positive, and FN is false negative. Average precision (AP) is the sum of all the precisions of a class in the validation set over the number of images with objects belonging to that class. AP was calculated to determine the mAP expressed as: $mAP = \frac{1}{11} \sum_{r\epsilon\{0.0,...,1.0\}} AP_r$, where mAP is the sum of the APs taken at 11 recall thresholds between 0 and 1.

The DSC (or F1 score) is given by the following expression: $DSC = \frac{2|M \cap A|}{|M|+|A|}$, where M is the area of expert manual segmentation, A is the area of the automated segmentation, and $M \cap A$ is the area of overlap between M and A. DSC has a range of (0,1) where the change from 0 to 1 represents no overlap to complete overlap.

The directed HD (h) between two point sets $A = \{a_1, ..., a_p\}$ the automated contour and $M = \{m_1, ..., m_{p'}\}$ the manual contour is the maximum distance between each point (p) from the automated contour to its nearest point (p') of the manual contour expressed as: $h(A, M) = \underset{p\epsilon A}{max}(\underset{p'\epsilon M}{min}\|p - p'\|)$, where $\|\cdot\|$ is any norm (e.g. the euclidean norm or L2). The HD is then the maximum directed HDs in both directions. HD between point sets A and M is given by: $HD(A, M) = max\left(h(A, M), h(M, A)\right)$.

3.4 Results and Discussion

3.4.1 Overall performance

Sample visual segmentation results for each of the eight datasets compared against the approximate ground truth are displayed in Fig. 3.6. The closely overlapping contours demonstrates that HMR-Net effectively classifies, locates and segments medical images in mul-

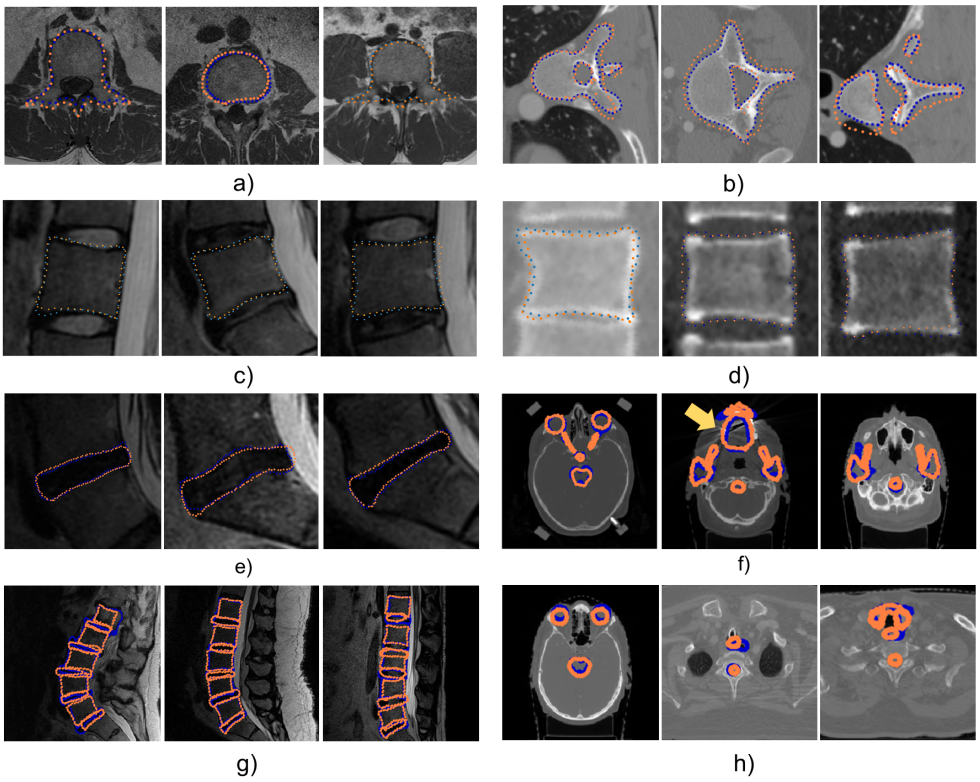


Figure 3.6. Sample HMR-Net segmentation results for each of the eight representative task datasets. a) Spine vertebra MR axial, b) spine vertebra CT axial, c) spine vertebral body MR sagittal, d) spine vertebral body CT sagittal, e) spine disc MR sagittal, f) head CT, g) spine MR sagittal, and h) head and neck CT. Manual delineations are in blue and HMR-Net's result is outlined in orange. HMR-Net is capable of segmenting organs in the presence of image artefacts as indicated by the yellow arrow.

tiple applications with great accuracy. Images from the eight tasks cover a broad range of medical applications, including a various organs spanning different regions of the body, displayed in different planes, and obtained from multiple imaging modalities. HMR-Net demonstrates excellent generality by overcoming the diversity of these images. HMR-Net achieves an accurate performance for organs of varying sizes, and is capable of capturing the ambiguous boundary of the oral cavity in the presence of image artefacts, indicated by the yellow arrow in Fig. 3.6(f). Quantitative results for each of the target organs among the eight datasets are presented in Table 3.4. DSC scores ranged from 0.44 for the thyroid to 0.91 for the oral cavity.

As shown in Fig. 3.6, the spine vertebra MR axial dataset consists of a single object, but with shape variations. Despite having a very limited number samples (50 images, refer to Table 3.1 for more details). HMR-Net exhibits a DSC score of 0.87 and a mAP of 0.75

Table 3.4. HMR-Net's performance for each target structure.

Structures	DSC	HD (mm)	$mAP_{50\text{-}95}$
Brainstem	0.81 ± 0.10	3.73 ± 0.65	0.51
Chiasm	0.48 ± 0.13	5.59 ± 0.58	0.33
Eye globe left	0.86 ± 0.14	2.79 ± 0.84	0.65
Eye globe right	0.84 ± 0.17	2.85 ± 0.86	0.62
Esophagus	0.85 ± 0.05	3.66 ± 0.57	0.59
Intervertebral discs	0.90 ± 0.05	3.89 ± 0.81	0.82
Larynx	0.84 ± 0.08	4.18 ± 0.75	0.57
Lips	0.73 ± 0.08	5.55 ± 0.55	0.73
Mandible	0.60 ± 0.20	3.57 ± 1.45	0.55
Optic nerve left	0.57 ± 0.21	3.90 ± 1.60	0.29
Optic nerve right	0.66 ± 0.19	3.70 ± 1.40	0.34
Oral cavity	0.91 ± 0.03	3.62 ± 0.83	0.72
Parotid left	0.67 ± 0.15	4.61 ± 1.05	0.36
Parotid right	0.64 ± 0.20	4.80 ± 1.00	0.33
Spinal cord	0.84 ± 0.06	2.94 ± 0.52	0.53
Submandibular left	0.78 ± 0.10	3.79 ± 0.68	0.43
Submandibular right	0.79 ± 0.10	3.58 ± 0.59	0.41
Thyroid	0.44 ± 0.22	3.72 ± 1.25	0.53
Vertebrae	0.86 ± 0.06	2.95 ± 0.49	0.83

Notation: Dice similarity coefficient (DSC) and Hausdorff distance (HD) are given in terms of mean (± standard deviation). The mean average precision (mAP) indicates the detection precision for classifying each organ. mAP reported here are at IoU thresholds from 50-95.

(see Table 3.5). This illustrates HMR-Net's robust capabilities even with small sample sets.

The spine vertebra CT axial dataset also consists of a single object, but contains immense geometric shape variations. Contrast to the vertebra MR axial dataset, the CT counterpart possesses more samples (170 images) for a single object task. However, the vertebra CT axial images contain much greater shape variations and complexity (e.g. objects with occlusions and multiple parts) (see Fig. 3.6). HMR-Net performs the worse for this task out of the other spine datasets with a DSC score of 0.75, but its performance is reasonable given the task's complexity.

Both the spine vertebral body and disc MR sagittal datasets represent cropped ROI versions of the spine MR sagittal dataset. The spine vertebral body CT sagittal dataset also repre-

sents a cropped ROI image task. HMR-Net achieves the highest scores for these three task datasets out of the total eight tasks with DSC scores of 0.93, 0.88, and 0.89 for the MR vertebral body, MR intervertebral disc, and CT vertebral body datasets respectively.

Spine MR sagittal, head CT, and the head and neck CT datasets represent a multiclass object segmentation task. The last column in Table 3.5 reports the performance of HMR-Net for these three datasets. For this complex task, HMR-Net achieved DSC scores of 0.85 for the spine MR sagittal multiclass dataset, 0.52 for head CT, and 0.66 for the head and neck CT dataset. For multiclass tasks, HMR-Net is capable of achieving a consistent high performance on medium to large size organs, illustrated by the spine MR dataset results (see Table 3.5).

However, the overall accuracy is dragged down when there are very small and complex organs present. A comparison between performing single class segmentation and multiclass segmentation for the head datasets validates this point (see Table 3.6). If a single class is only considered HMR-Net can achieve an average DSC accuracy of approximately 0.70 across all the organs in head and neck region, but when learning all the classes simultaneously the accuracy drops by roughly 10%. This drop in performance indicates a limitation in the object detection branch for handling imbalanced samples for small and complex organ structures (organs with large geometric variability). Nevertheless, this performance drop was well within the expectations for the multiclass object segmentation task performed in this study. HMR-Net's ability to perform multiclass object segmentation is an additional feature to further demonstrate the network's generalizability for multiple applications.

3.4.2 Ablation experiments

Ablation experiments were performed to demonstrate the effectiveness of the manifold regularization loss and the employment of cross-stitch units. As shown in Table 3.5, HMR-Net's performance including both components achieves the highest score for a majority of the tasks indicating the benefit of incorporating them. The absence of either component resulted in a reduction in performance.

For instance, the spine MR axial dataset's DSC score improved approximately 7-9% from 0.80 and 0.78 to 0.87 when both the cross-stitch layers and manifold regularization components were used. mAP scores also improved from 0.67 and 0.70 to 0.75. The absence of the manifold regularization provided the greatest reduction in performance indicating

Table 3.5. HMR-Net's ablation study results.

Condition	Manifold Regularization		Cross-stitch Layers		HMR-Net (All)	
Dataset	DSC	$mAP_{50\text{-}95}$	DSC	$mAP_{50\text{-}95}$	DSC	$mAP_{50\text{-}95}$
SD-MR-Sag	0.82 ± 0.07	0.70	**0.91 ± 0.05**	**0.85**	0.91 ± 0.06	0.74
SV-MR-Sag	0.90 ± 0.02	0.74	**0.94 ± 0.02**	**0.85**	0.93 ± 0.03	0.81
SV-MR-Ax	0.78 ± 0.03	0.67	0.80 ± 0.10	0.70	**0.87 ± 0.03**	**0.75**
SV-CT-Sag	0.87 ± 0.06	0.80	**0.92 ± 0.03**	**0.87**	0.89 ± 0.06	0.77
SV-CT-Ax	0.69 ± 0.18	0.79	0.67 ± 0.20	0.67	**0.75 ± 0.11**	**0.77**
S-MR-Sag	0.82 ± 0.06	0.71	0.84 ± 0.06	0.71	**0.85 ± 0.10**	**0.74**
Head (multiclass)	0.44 ± 0.20	0.32	0.50 ± 0.19	0.35	**0.52 ± 0.19**	**0.37**
HaN (multiclass)	0.54 ± 0.21	0.10	0.54 ± 0.21	0.20	**0.66 ± 0.15**	**0.26**

Notation: Dice similarity coefficient (DSC) and Hausdorff distance (HD) are given in terms of mean (± standard deviation). Mean average precision (mAP) at IoU thresholds from 50-95. The highest DSC and mAP values are in bold.

the importance of preserving a data's local geometric structure. This can be seen by the difference in results from all eight datasets especially from the head and neck multiclass dataset with improved DSC scores of 0.54 to 0.66 and a mAP of 0.10 to 0.26.

3.4.3 Literature comparative analysis

A comparative study against current state-of-the-art algorithms was conducted comparing HMR-Net against the widely used U-Net [17] as a general framework for medical image segmentation, and RegressionCNN [25], a shape regression framework. As shown in Table 3.6 column 1, HMR-Net is comparable or superior to both U-Net and RegressionCNN for each of the eight representative task datasets in terms of overall segmentation overlap.

However, the HD that represents the maximum distance between each point from the predicted contour to its nearest point of the manual contour was best minimized by RegressionCNN (as shown in Table 3.6, column 2). This degrade in performance for HMR-Net contrast to RegressionCNN can be attributed to the effect of negative transfer from the detection network branch in multitask networks (described in Section 3.2.2). When performing joint organ detection and segmentation, the segmentation results are heavily dependent on the detection branch's performance. The positioning of the predicted bounding box can offset and affect the regression segmentation's predicted position as well. Extracted features for the coarse-to-fine localization of the target organ does not provide enough precision to

Table 3.6. State-of-the-art model comparisons for the eight applications.

	U-Net		Regression CNN		HMR-Net	
Datasets	DSC	HD (mm)	DSC	HD (mm)	DSC	HD (mm)
SD-MR-Sag	0.85 ± 0.10	2.74 ± 0.58	0.75 ± 0.11	**1.88 ± 0.33**	**0.91 ± 0.06**	3.89 ± 0.88
SV-MR-Sag	0.89 ± 0.07	2.75 ± 0.55	0.86 ± 0.06	**1.80 ± 0.35**	**0.93 ± 0.03**	4.07 ± 0.97
SV-MR-Ax	0.48 ± 0.15	2.37 ± 0.60	0.81 ± 0.05	**1.69 ± 0.18**	**0.87 ± 0.03**	2.70 ± 0.43
SV-CT-Sag	0.79 ± 0.15	3.54 ± 0.54	0.80 ± 0.05	**1.68 ± 0.13**	**0.89 ± 0.06**	3.64 ± 0.43
SV-CT-Ax	0.60 ± 0.21	2.28 ± 0.53	0.58 ± 0.18	**1.73 ± 0.27**	**0.75 ± 0.11**	2.23 ± 0.29
Head (single-class)	0.51 ± 0.18	3.86 ± 0.69	0.67 ± 0.14	**3.20 ± 0.54**	**0.70 ± 0.11**	4.31 ± 1.15
HaN (single-class)	0.55 ± 0.19	3.72 ± 0.77	**0.73 ± 0.09**	**3.20 ± 0.55**	0.71 ± 0.08	3.40 ± 0.63
S-MR-Sag	N/A	N/A	N/A	N/A	**0.85 ± 0.10**	2.18 ± 0.34
Head (multiclass)	N/A	N/A	N/A	N/A	**0.52 ± 0.19**	3.49 ± 0.67
HaN (multiclass)	N/A	N/A	N/A	N/A	**0.66 ± 0.15**	3.43 ± 0.78

Notation: Dice similarity coefficient (DSC) and Hausdorff distance (HD) are given in terms of mean (± standard deviation). The highest DSC values and the lowest HD values are in bold.

aid the shape regression task and may potentially lead some points astray. Although we implemented cross-stitch layers to attempt to remedy the negative transfer challenge, it is possible that allowing the network to learn similarly to sluice networks [35] with a beta parameter to control the output layers for each task may help improve the overall final results even further.

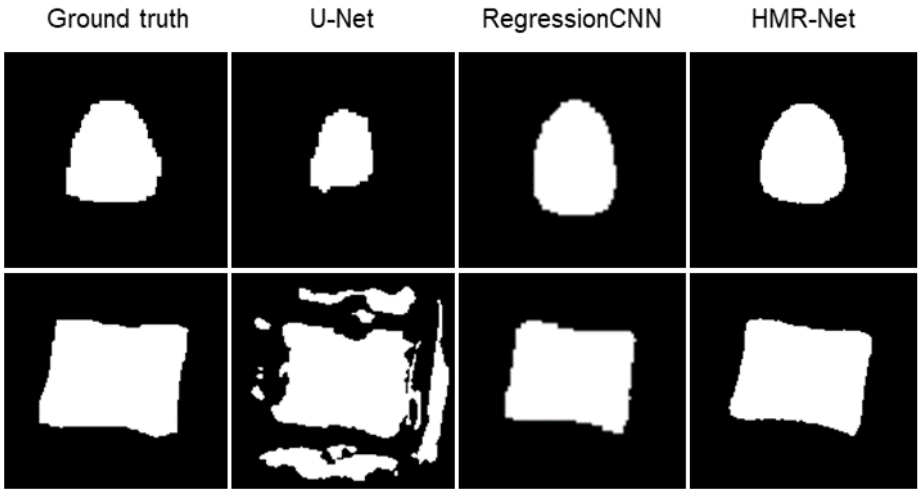


Figure 3.7. Sample visual results from the state-of-the-art models compared with HMR-Net. HMR-Net's result, depicted in the far right, is the most comparable to the approximate ground truth in contrast to the other two methods.

U-Net was designed best for binary segmentation tasks, thus for our multiclass datasets we trained both U-Net and HMR-Net for each organ class as a binary segmentation problem. RegressionCNN was designed as a single class shape regression algorithm with its target organs centred and cropped. The training protocol for both U-Net and RegressionCNN followed their respective literature descriptions to achieve optimal results. The binary cross-entropy loss was used for U-Net and the MSE loss was used for RegressionCNN.

Sample results from each comparison algorithm are displayed in Fig. 3.7. HMR-Net's segmentation results are the most comparable to the approximate ground truth compared to both U-Net and RegressionCNN. Our proposed framework achieves the best performance for a majority of the task datasets examined in this paper (see Table 3.6). HMR-Net even outperforms the semantic segmentation-based model U-Net.

3.4.4 Potential applications

The proposed method holds great potential for the following applications:

1) Image analysis applications. Directly, HMR-Net was trained on eight representative task datasets that demonstrate its potential performance for the following: single object and multi-object segmentation, patch images and full-scale images, large and small datasets, balanced and imbalanced datasets, image slices on different planes and imaging modalities, and datasets with few to large object size variability.

2) Post-processing applications. HMR-Net tags and identifies multiple different objects along with their location and exact boundaries, enabling easier post-information processing and association. Whereas in clinical routine physicians have to manually record organ labels and documentation details. The proposed algorithm can automatically extract and provide this information helping to reduce one more tedious task in the radiology diagnostic workflow.

Other potential applications for the proposed network are not limited to the list above, but can be extended to other organs in the body such as cardiac structures, liver, kidney, knee, clavicle, among others.

3.4.5 Highlights of the HMR-Net architecture

The proposed HMR-Net architecture performance benefits from the following aspects:

1) The framework considers the relationship and learning behaviour of loosely related tasks and rebalances the learned features for detection and segmentation with a trainable alpha parameter. The weights are redistributed to better fit the two specific tasks during training.

2) The feature pyramid network provides multiscale feature maps from different layers of the network best suited for different object sizes. Small objects require more local and structural features that are obtained in the lower levels of the network. Whereas large objects benefit more from the higher-level features with greater semantic information.

3) The bottom-up pathway after the feature pyramid network was constructed to propagate strong responses of low-level patterns to top-level layers necessary to accurately locate object instances. The higher resolution feature map outputs, generated from the bottom-up pathway, then helps to enhance the retention of lower-level features to solve the challenging regression tasks.

4) Merging of the feature map outputs only for shape regression enhanced the point estimation task by providing both low-level and top-level features necessary for finer precision.

As an additional option, the two branch networks can also function separately allowing for only the detection or segmentation branch to function on their own or together.

3.4.6 Limitations

The main limitations of the proposed framework are: 1) small and complex organs still remain difficult to obtain precise detection and segmentation results, 2) an average or mean organ shape is generated in samples belonging to the same class but possess various shapes. This occurs when there is an imbalance in samples across the various shapes. 3) The segmentation results rely on the accuracy of the detection branch. False positives and false negatives will produce extra or missing segmentation results, respectively.

References

[1] B. I. Reiner, E. Siegel, and K. Siddiqui, "Evolution of the digital revolution: A radiologist perspective," *Journal of Digital Imaging: the Official Journal of the Society for Computer Applications in Radiology*, vol. 16, pp. 324–330, 2004.

[2] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Medical Physics*, vol. 44, no. 2, pp. 547–557, 2017.

[3] V. Pekar, S. Allaire, A. Qazi, J. Kim, and D. Jaffray, "Head and neck auto-segmentation challenge: Segmentation of the parotid glands," *MICCAI 2010: A Grand Challenge for the Clinic*, pp. 273–280, 2010.

[4] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, B. M. Dawant, *et al.*, "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015," *Medical Physics*, vol. 44, no. 5, pp. 2020–2036, 2017.

[5] J. Karsten and O. Arandjelović, "Automatic vertebrae localization from CT scans using volumetric descriptors," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 576–579, 2017.

[6] A. B. Oktay and Y. S. Akgul, "Simultaneous localization of lumbar vertebrae and intervertebral discs with SVM-based MRF," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 9, pp. 2375–2383, 2013.

[7] 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.

[8] S. Ghosh, M. R. Malgireddy, V. Chaudhary, and G. Dhillon, "A new approach to automatic disc localization in clinical lumbar MRI: Combining machine learning with heuristics," *Proceedings - International Symposium on Biomedical Imaging*, pp. 114–117, 2012.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 1, pp. 886–893, 2005.

[10] K. Kim and S. Lee, "Vertebrae localization in CT using both local and global symmetry features," *Computerized Medical Imaging and Graphics*, vol. 58, pp. 45–55, 2017.

[11] S. Gregory, F. K. D., P. Vladimir, P. Marta, S. Nadya, V. Harini, *et al.*, "Vision 20/20: Perspectives on automated image segmentation for radiotherapy," *Medical Physics*, vol. 41, no. 5, p. 50902, 2014.

[12] J.-F. Daisne and A. Blumhofer, "Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation," *Radiation Oncology*, vol. 8, no. 1, p. 154, 2013.

[13] V. Pekar, S. Allaire, J. Kim, and D. A. Jaffray, "Head and Neck Auto-segmentation Challenge," *Midas Journal*, pp. 349–445, 2009.

[14] A. Qazi, V. Pekar, J. Kim, J. Xie, S. Breen, and D. Jaffray, "Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach," *Medical Physics*, vol. 38, pp. 6160–6170, 2011.

[15] P. Boonyakiat and P. Silapachote, "Segmentation of optic nerve head images," in *2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–5, 2017.

[16] I. Ben Ayed, K. Punithakumar, G. Garvin, W. Romano, and S. Li, "Graph cuts with invariant object-interaction priors: Application to intervertebral disc segmentation," in *Information Processing in Medical Imaging* (G. Székely and H. K. Hahn, eds.), (Berlin, Heidelberg), pp. 221–232, Springer Berlin Heidelberg, 2011.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *arXiv e-prints arXiv:1505.04597*, 2015.

[18] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *ArXiv*, vol. abs/1802.06955, 2018.

[19] B. Zhao, J. Soraghan, G. D. Caterina, and D. Grose, "Segmentation of head and neck tumours using modified U-net," in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–4, 2019.

[20] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in CT images," *Medical Image Analysis*, vol. 13, no. 3, pp. 471–482, 2009.

[21] S. H. Huang, Y. H. Chu, S. H. Lai, and C. L. Novak, "Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI," *IEEE Transactions on Medical Imaging*, vol. 28, no. 10, pp. 1595–1605, 2009.

[22] X. Zhu, X. He, P. Wang, Q. He, D. Gao, J. Cheng, *et al.*, "A method of localization and segmentation of intervertebral discs in spine MRI based on gabor filter bank," *Biomedical Engineering Online*, vol. 15, no. 1, p. 32, 2016.

[23] Y. Wang, L. Zhao, M. Wang, Z. Song, and M. Wang, "Organ at risk segmentation in head and neck CT images by using a two-stage segmentation framework based on 3D U-Net," *arXiv preprint arXiv:1809.00960*, no. 2, pp. 1–11, 2018.

[24] Z. Wang, X. Zhen, K. Tay, S. Osman, W. Romano, and S. Li, "Regression segmentation for M^3 spinal images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 8, pp. 1640–1648, 2015.

[25] J. Chen, H. Zhang, W. Zhang, X. Du, Y. Zhang, and S. Li, "Correlated regression feature learning for automated right ventricle segmentation," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–10, 2018.

[26] X. Zhen, Y. Yin, M. Bhaduri, I. B. Nachum, D. Laidley, and S. Li, "Multi-task shape regression for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 210–218, Springer, 2016.

[27] T. He, J. Guo, J. Wang, X. Xu, and Z. Yi, "Multi-task learning for the segmentation of thoracic organs at risk in CT images," in *SegTHOR@ISBI*, pp. 10–13, 2019.

[28] R. Ke, A. Bugeau, N. Papadakis, P. Schuetz, and C.-B. Schönlieb, "A multi-task U-net for segmentation with lazy labels," *arXiv e-prints*, p. arXiv:1906.12177, jun 2019.

[29] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR: Clinical Orthopaedics and Related Research*, vol. abs/1703.0, 2017.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[31] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR: Clinical Orthopaedics and Related Research*, vol. abs/1612.0, 2016.

[32] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," *CoRR: Clinical Orthopaedics and Related Research*, vol. abs/1803.0, 2018.

[33] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the CoordConv solution," *32nd Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1–26, 2018.

[34] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," 2016.

[35] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Latent multi-task architecture learning," *arXiv preprint arXiv:1705.08142*, 2017.

[36] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," 2013.

[37] X. Du, W. Zhang, H. Zhang, J. Chen, Y. Zhang, J. Claude Warrington, *et al.*, "Deep regression segmentation for cardiac Bi-Ventricle MR images," *IEEE Access*, vol. 6, pp. 3828–3838, 2018.

[38] C. Tam, X. Yang, S. Tian, X. Jiang, J. Beitler, and S. Li, "Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression," *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 10578, 2018.

[39] S. Pang, Z. Su, S. Leung, I. B. Nachum, B. Chen, Q. Feng, *et al.*, "Direct automated quantitative measurement of spine by cascade amplifier regression network with manifold regularization," *Medical Image Analysis*, vol. 55, pp. 103–115, 2019.

[40] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[41] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 1225–1234, ACM, 2016.

[42] F. C. Graham and F. R. K. Chung, *Spectral Graph Theory*. Regional conference series in mathematics, American Mathematical Society, 1996.

[43] S. Butler, "Generalizing some results to the normalized Laplacian," pp. 1–11, 2006.

[44] 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.

[45] F. Chollet, "Keras." https://keras.io, 2015.

[46] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016.

[47] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow." https://github.com/matterport/Mask_RCNN, 2017.

[48] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," *Encyclopedia of Database Systems. Springer, Boston, MA*, 2009.

CHAPTER 4

The concluding chapter of this thesis revisits the motivation, research objectives, and summarizes the important findings and conclusions of Chapters 2 and 3. The limitations and future studies regarding machine learning technology in medical image segmentation inspired by the work presented in this thesis will be discussed along with its impact to the field.

4 CONCLUSION AND FUTURE DIRECTIONS

4.1 Overview of Rationale and Research Questions

In a technology-oriented field, such as image analysis, there is a constant flow of innovative technology coming and going. Where new methods presented within the last two years are now considered old, but in actuality have only just begun to be validated by the research and scientific community. Machine learning and deep learning-based algorithms, in particular for image segmentation, have evolved substantially to provide a high-level of understanding of an image's semantic context and properties. Although the developments of new machine learning and deep learning techniques are produced every year, many of them are specialized for a single application. The practical use for these types of non-reusable algorithms are not optimal in clinical routine.

The goal of designing automated image processing algorithms is to serve as an ancillary service for image analysis. One of the targeted applications for machine learning implementation is for image segmentation. Segmentation is the principal and initial step preceding other image analysis tasks and to diagnostic and therapeutic procedures. However, due to the current task specific limitations of most semi-automated and automated segmentation algorithms manual delineations is still the standard clinical routine.

The design of a general automated segmentation tool would greatly provide a standardized method for image segmentation practices, and aid to eliminate interobserver variability that comes with manual segmentation. Accumulated knowledge from several physicians could be stored in this general segmentation system.

However, the scope in designing a general segmentation model is large and complex, and the techniques involved to create one are not greatly explored. Therefore, in order to work towards this goal, the overarching theme of this thesis was to leverage machine learning

and deep learning's technological advancements to design a multiapplication segmentation tool. The specific research objectives that were conducted here include: 1) exploring a novel multitask regression approach to image segmentation that formulates the task into a shape or boundary regression problem (Chapter 2). 2) Combine the concept of multitask learning and multitask regression to generate a multiapplication segmentation framework (Chapter 3).

4.2 Summary and Conclusions

In Chapter 2, we proposed a multi-output support vector regression (MSVR) model that accurately and efficiently segmented the organs-at-risk (OAR) using a representative database of head and neck CT images from 56 subjects. Segmentation results of the OAR was demonstrated to be similar or superior to baseline results of state-of-the-art algorithm CNN [1]. The processing time of our method demonstrated the effectiveness of machine learning capabilities in providing efficient segmentation results (See Chapter 2 Section 3.1). Limitations for smaller low contrasting organs or organs with great shape variance can be improved by integrating other feature descriptors to capture more discriminant information, such as in Wang et al. [2] and He et al. [3]. However, based on current results against manual segmentation along with the literature comparison, our model still demonstrated its capabilities in providing robust, accurate, and efficient predictions.

In Chapter 3, we proposed a multiapplication framework for image segmentation as a step towards standardizing segmentation practices. The proposed holistic multitask regression network (HMR-Net) accurately and efficiently classified, located, and segmented all 19 organs in the eight representative task datasets from 222 subjects (53 head, 56 head and neck, 113 spine). Our model demonstrated its capabilities as an effective framework by providing efficients and accurate predictions by achieving mAP and DSC scores reaching up to 0.81 and 0.93, respectively. HMR-Net successfully handles multiplane, multimodality, multistructural, and multiregional images with a single framework. HMR-Net performs accurate segmentation for a variety of structures with a wide range in sizes, which significantly represents HMR-Net's ability to successfully handle various complex organ shapes and sizes. Thus, our model holds exciting potential as an ancillary application for preliminary image processing and analysis. Future work aims to further develop HMR-Net's architecture to for 3D image applications, as well as an automated adaptive hyperparameter optimization pre-processing algorithm.

4.3 Limitations

4.3.1 Study specific limitations

Chapter 2: Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector

In Chapter 2, the proposed MSVR algorithm obtained lower dice scores for smaller low contrasting organs or organs with great shape variance that suggested limitations in the proposed algorithm. The MSVR presented in this chapter used a single feature descriptor in order to represent the input image data, which was the histogram of oriented gradients. With this one descriptor alone, segmentation of medium to larger sized organs achieved a dice score of over 80%. However, a single descriptor detailing the changes in gradient values would not be sufficient to capture ambiguous boundaries not supported by a homogenous background. Contours for these small low contrast organs were still recoverable due the holistic boundary regression technique, preventing any of the organs to achieve a dice score of less that 60%. However, as discussed in Chapter 2, supplementary information can be provided by multiple feature descriptors to overcome this limitation. Automated feature descriptors, such as deep learning, would allow the algorithm to learn which features to extract from the image; thus, removing the challenge of deciding which additional descriptor to use.

Furthermore, another limitation in this study is that only a single object or organ can be processed at a time. Although the proposed method can be applied to multiple organs, this method is only capable of achieving optimal results on patched or cropped ROI images. This limits the scope of applications. Processing a single organ at a time is still time consuming and not ideal, despite reducing some of the processing time by semi-automation. Incorporating an automated ROI detection algorithm would remove the initial manual preprocessing and allow for fully automated processing.

Chapter 3: Holistic multitask regression network for multiapplication shape regression segmentation

In Chapter 3, the proposed deep learning network, HMR-Net, was limited for detecting extremely small organs in datasets with large organ size discrepancies along with imbalanced samples for multiclass object detection. An attention type mechanism could provide more constraints to limit the area of prediction for extremely small organs, thus improving the

detection precision. Although the imbalanced samples challenge was addressed by generating augmented samples for the rare classes, it was not sufficient enough to fully solve the problem. For the imbalance of samples, giving a weight to the classes based on their frequency can help to further reduce misdetections of the rare class samples.

For small datasets with great organ shape diversity, only by providing more data can the network learn all the variations of that organ of interest. However, the training data does not necessarily have to be all labelled data. Unsupervised learning or semi-supervised learning techniques can be adopted to overcome the limitation of insufficient labelled samples.

4.3.2 General limitations

The development and validation of machine learning and deep learning algorithms in the field possess a set of limitations that also needs to be addressed. One such limitation is data bias. Many algorithms if not carefully validated are biased towards the training data used to develop it. Even with the use of cross-validation methods, there still exists a slight positive bias towards the developer's dataset. This is because decisions on how to improve the network would be based on the test samples from the developer's overall dataset. Unless a separate dataset is completely left out for validation purposes and not used to influence the development of the machine learning algorithm, a dataset bias would be present.

However, having enough data to split your dataset into three cannot always be done. The disparity in the field of medical image analysis is the lack of available data, or the difficulty in obtaining the required data for a developer's specific task. For computer vision obtaining natural images is not nearly as restricting as it is in the medical domain. Patient privacy must be adhered to, and images obtained from medical imaging devices are costly for each scan. Medical images can only be obtained from medical professionals and not self acquired. These data resource limitations is an area that requires the combined cooperation of medical experts, patients, and scientific researchers to resolve.

4.4 Future Directions

4.4.1 Automated detection of small organs in imbalanced datasets

For future studies, the results in Chapter 3 has prompted the re-examination of HMR-Net's classification and localization branch. The challenges to detect small organs and to over-

come imbalanced datasets are still on-going problems in the field with a few solutions that tackle it to some degree. The exploration of integrating of other detection methods, compared to our two-stage candidate box proposal technique, would provide more insight into resolving the current detection performance limitation.

4.4.2 Task adaptive hyperparameter optimization

The long training process of manual hyperparameter tuning is another topic that has sparked a lot interest. Although there are current techniques to optimize the hyperparameters to a deep learning network, such as grid search [4] and random search [5], they are not suitable for complex frameworks and are computationally heavy. Grid search requires some previous knowledge to limit the combinations of hyperparameters to optimize. Whereas random search can optimize the best combination for all the hyperparameters. The development of a task adaptive hyperparameter optimization method is another field a research worth exploring.

4.4.3 Three-dimensional (3D) image segmentation

To further develop our multiapplication framework into a general segmentation tool we seek to extend its capabilities to 3D image applications. Since our method performs shape regression with x and y coordinates, the addition of a z dimension would not be a huge modification to the current network structure. In addition, the 2D convolutional layers would simply need to be replaced with 3D convolutional layers. 3D image segmentation can provide more information into an organ's geometry and is favoured compared to its 2D counterpart. However, processing 3D images will be computationally heavy, and it is something we will need to address with this modification. With the addition of 3D images, if successful, this would bring our model that much closer to a general segmentation tool.

4.4.4 Insight into the information sharing process of multitask learning networks

Lastly, looking into the behaviour of the alpha parameters in cross-stitch units would help to provide more insight into the learning process of deep learning networks performing multitask learning. The amount of information shared among the different tasks in a multitask network, if automatically determined by a deep learning network during training, is

not yet fully understood. By exploring this process it would give us a better understanding of how deep learning algorithms interact with different tasks.

4.5 Significance and Impact

While there are many machine learning and deep learning algorithms with sufficient performance in accuracy, reliability, and speed, the current field requires more generic image analysis technology that can be effectively adapted for a specific clinical task. Methods that are reliable, fast, accurate, and generalizable are highly sought for. Deep learning model combinations utilizing multitask learning has provided the machine learning field with a solution to create more generalized networks.

This thesis has increased the fundamental knowledge of multitask learning networks and presented an innovative approach to segmentation, shape regression. The shape regression technique was developed solely within our lab. From the studies presented in this thesis, we have presented the first multiapplication deep learning segmentation framework for multi-object shape regression. This development brings us much closer to achieving a general segmentation tool beneficial to clinical routine.

References

[1] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Medical Physics*, vol. 44, no. 2, pp. 547–557, 2017.

[2] Z. Wang, X. Zhen, K. Tay, S. Osman, W. Romano, and S. Li, "Regression segmentation for M^3 spinal images," *IEEE Transactions on Medical Imaging*, vol. 34, no. 8, pp. 1640–1648, 2015.

[3] X. He, A. Lum, M. Sharma, G. Brahm, A. Mercado, and S. Li, "Automated segmentation and area estimation of neural foramina with boundary regression model," *Pattern Recognition*, vol. 63, pp. 625–641, 2017.

[4] R. Ghawi and J. Pfeffer, "Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity," *Open Computer Science*, vol. 9, no. 1, pp. 160–180, 2019.

[5] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

APPENDIX

APPENDIX A: Reprint Permissions for Scientific Article

Curriculum Vitae

Name: Clara M.Tam

Post-Secondary Education and Degrees:

The University of Western Ontario
London, ON, Canada
Medical Biophysics
2014-2017 Bachelor of Medical Science

The University of Western Ontario
London, ON, Canada
Biomedical Engineering, Imaging in Musculoskeletal Health Research
2017-2020 Masters in Engineering Science

Honours and Awards:

Western Graduate Research Scholarship
Institutional
2017-2019

Transdisciplinary Bone & Joint Training Award
Collaborative Training Program in Musculoskeletal Health Research
Institutional
2017-2019

Related Work Experience:

Undergraduate Research Trainee
The University of Western Ontario
London, ON, Canada
2016-2017

Research Assistant
The University of Western Ontario
London, ON, Canada
2017

Graduate Student Research Assistant, Masters
The University of Western Ontario
London, ON, Canada
2017-2020

Publications and Presentations

A. Refereed Journal Manuscripts (1 Published, 2 In preparation)

Published (1)

1. Luo G, Dong S, Wang W, Wang K, Cao S, **Tam C**, Zhang H, Howey J, Ohorodnyk P, and Li S. (2020). Commensal correlation network between segmentation and direct area estimation for bi-ventricle quantification. Med. Image Anal., vol. 59, p. 101591. (2020/1).

In Preparation (2)

1. Dong S, Luo G, **Tam C**, Wang W, Wang K, Cao S, Chen B, Zhang H, and Li S. (2019). Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. Med. Image Anal. (Under revisions; 2019/12)

2. **Tam C**, Zhang D, Peters T, and Li S. (2019). Holistic multitask regression network for multiapplication shape regression segmentation. Med. Image Anal. (Submitted; 2019/12)

B. Published Refereed Conference Paper (1)

1. **Tam C**, Yang X, Tian S, Jiang X, Beitler JJ, and Li S. (2018). Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression. Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging. SPIE Medical Imaging Conference.

C. Published Refereed Conference Abstracts (2)

1. **Tam C**, He X, and Li S. (2017). Boundary regression segmentation of spinal images. 15th Annual Meeting of Imaging Network Ontario Program. 15th Annual Imaging Network Ontario (IMNO) Symposium. (2017/3)

2. **Tam C**, He X, Sharma M, Mercado A, Landis M, and Li S. (2017). Automated multi-vertebrae and disc delineation for MR and CT spinal images. Explore. Invent. Transform. RSNA 2017. Radiological Society of North America (RSNA). (2017/11)

D. Presentations (4)

1. **Tam C**, He X, and Li S. (2017). Boundary regression segmentation of spinal images. 15th Annual Meeting of Imaging Network Ontario Program. 15th Annual Imaging Network Ontario (IMNO) Symposium IMNO Symposium. Poster Presentation (2017/3)

2. **Tam C**, He X, Sharma M, Mercado A, Landis M, Li S. (2017). Automated multi-

vertebrae and disc delineation for MR and CT spinal images. Explore. Invent. Transform. RSNA 2017. Radiological Society of North America (RSNA). Poster Presentation (2017/11)

3. **Tam C**, Yang X, Tian S, Jiang X, Beitler JJ, Li S. (2018). Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression. Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging. SPIE Medical Imaging Conference. Poster Presentation (2018/2)

4. **Tam C**, Peters T, and Li S. (2019). AI's got your back. 3-Minute Thesis. Innovation in Motion: The Faculty of Health Sciences and School of Biomedical Engineering Research Day Program. Oral Presentation (2019/6)