University of London
Imperial College of Science, Technology and Medicine
Department of Computing

Machine Learning for Analysing Whole-body Scans

Vanya Vabrina Valindria

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London

December 2018

Abstract

Advances in machine learning techniques have been shown to bring benefit for analysing medical images. Whole-body scans contain multiple organs, which makes the manual annotation harder, hence the amount of data obtained is still limited. Meanwhile, machine learning needs a large amount of annotated data in order to perform well.

Image segmentation is a crucial task in medical image analysis and beneficial for clinical diagnosis. One of the segmentation challenges is to infer the automatic segmentation quality between algorithms. Unfortunately, the manual annotations required to compute the segmentation accuracy are sometimes unavailable. In order to tackle this, we propose a new method to predict the segmentation evaluation on a per-case basis without annotated data.

Segmenting scans acquired from different imaging centers than the one used for algorithm training, can lead to degraded segmentation results. This problem has brought the necessity of domain adaptation. We studied how to select useful samples from unlabeled target domain for supervised domain adaptation. This framework is shown to be less time-consuming with promising results to be useful in clinical practice.

A novel dual-stream network is also introduced to learn useful features on unpaired images from different modalities. This network demonstrates its power in multi-modal learning to segment multiple abdominal organs on both modalities. Additionally, strategies to segment challenging small organs on whole-body scans are investigated. We show that using two-stage networks with weighting schemes is useful to solve class imbalance and multi-scale contextual information.

Machine learning is not only helping the work of radiologists with faster automatic segmentation tools, but also leveraging the useful features for alleviating real clinical problems. We believe that our proposed methods could offer an important contribution to the field of medical image analysis; advancing our understanding on how machine-learning approaches could be applied in several problems in whole-body image analysis.

Acknowledgements

First and foremost I wish to thank my supervisor Prof. Daniel Rueckert and Dr. Ben Glocker, for giving me the opportunity to pursue a PhD degree. I am very grateful to have supportive supervisors like them, who are very patiently guide me throughout this PhD journey. They understand how challenging it is to juggle between PhD, a toddler and a baby; with other random stuff.

I acknowledge BPRI (Indonesian Presidential Scholarship) via LPDP (Indonesian Endowment Fund for Education) that has been sponsored my PhD study for 4 years.

I would like to thank to people at the BioMedIA group for giving me such a good atmosphere to learn in research environment. Special thanks for Wenjia, Kostas, Martin, and Nick, Juan, for their valuable input and help to complete the research work presented in this thesis. Thank you to Benjamin and Rob who helped me in English proof-reading for this thesis, also to Amir who always supplies me with updated info during this PhD journey.

I also thank Ioanis Lavdas, Eric O. Aboagye, Andrea G. Rockall for their MALIBO data (plus annotation) and clinical feedback. Then, I would like to thank the examiners: Marleen de Bruijne and Abhijeet Ghosh for their time and willingness to read and to correct this thesis.

I am extremely grateful to my parents, for their constant prayers and support. Finally, I thank my husband and children - who have gone miles and miles away to be always by my side to give encouragement.

Declaration of originality

I, Vanya Valindria, hereby declare that the work described in this thesis is my own, except where specifically acknowledged.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

To my Rhapsody and Symphony...

Acronyms

RCA Reverse Classification Accuracy

MRI Magnetic Resonance Imaging

RF Radio Frequency

CT Computed Tomography

MALIBO MAchine Learning In whole Body Oncology

GT ground truth

CNN Convolutional Neural Networks

ROI Region of Interest

MSL Marginal Space Learning

SVM Support Vector Machine

MMD Maximum Mean Discrepancy

GAN Generative Adversarial Network

NAF Neighbourhood Approximation Forest

WESD Weighted Spectral Distance for Measuring Shape Dissimilarity

CBIR Content-Based Image Retrieval

MI Mutual Information

t-SNE t-Distributed Stochastic Neighbor Embedding

PCA Principle Component Analysis

MDS Multi Dimensional Scaling

LLE Locally Linear Embedding

DSC Dice's similarity coefficient

FCN Fully convolutional network

Contents

Abstract **i**

Acknowledgements **iii**

Acronyms **viii**

1 Introduction **1**

1.1 Background and Motivation 1

1.1.1 The Anatomy of Human Body 2

1.1.2 Magnetic Resonance Imaging (MRI) 4

1.1.3 Computed Tomography (CT) 7

1.2 Datasets Used 8

1.2.1 MRI Whole-body Images from MALIBO dataset 8

1.2.2 MRI Whole-body Images from UK Biobank dataset 9

1.2.3 CT Abdominal Images 10

1.3 Technical Challenges in Whole-body Image Analysis 10

1.4 Thesis Contributions 12

1.5 Thesis Overview 13

1.6 List of Publications 14

2 Review of Machine Learning Methods in Whole-Body Image Analysis 16

2.1 Whole-Body Image Segmentation 17

2.1.1 Multi-atlas Segmentation 18

2.1.2 Random Forest 20

2.1.3 Convolutional Neural Networks (CNN) 26

2.2 Transfer Learning 31

2.2.1 Supervised Domain Adaptation 32

2.2.2 Semi-supervised Domain Adaptation 33

2.2.3 Unsupervised Domain Adaptation 34

2.3 Multi-modal Learning 36

2.3.1 Learning on Paired Images 37

2.3.2 Learning on Unpaired Images 37

2.4 Population Image Analysis 38

2.4.1 Quality Control 39

2.4.2 Manifold Learning for Population Study 45

3 Learning to Predict the Segmentation Accuracy 48

3.1 Introduction 49

3.1.1 Related Work 50

3.1.2 Contribution 52

3.2 Reverse Classification Accuracy (RCA) 53

3.2.1 Learning Reverse Classifiers 54

3.2.2 Predicting Segmentation Accuracy 56

3.2.3 Summary 58

3.3 Experiments and Results 59

3.3.1 Experimental Setting 59

3.3.2 Quantifying Prediction Accuracy using Different Reverse Classifiers . . . 60

3.3.3 Results for Predicting Dice's Similarity Coefficients 60

3.3.4 Predicting Degraded Segmentation Accuracies 65

3.3.5 Results for Predicting Different Segmentation Metrics 66

3.4 Discussion and Conclusion 69

4 Learning between Different Domains 73

4.1 Introduction 74

4.2 Methodology 75

4.2.1 Datasets 75

4.2.2 Domain Adaptation using Reverse Classification Accuracy (DARCA) 77

4.2.3 Supervised Domain Adaptation 77

4.3 Experiments and Results 78

4.3.1 Training from Scratch 79

4.3.2 Fine-tuning a Pre-trained Network 80

4.3.3 Iterative DARCA 82

4.3.4 DARCA on Right Kidney Segmentation 82

4.4 Discussion and Conclusion 86

5 Learning between Imaging Modalities 87

5.1 Introduction 88

5.2 Methodology 91

5.2.1 Network architectures 91

5.2.2 Dual-stream 93

5.2.3 Dual-stream Networks for Multi-modal learning on unpaired images . . . 93

5.3 Experiments and Results 94

5.3.1 FCN-based dual-stream Networks 95

5.3.2 U-Net based Dual-stream Networks 104

5.4 Discussion and Conclusion 108

6 Learning to Segment Small Organs 110

6.1 Introduction 111

6.2 Methodology 114

6.2.1 Materials 114

6.2.2 Two-stage network: A coarse-to-fine approach 116

6.3 Experiments and Results 118

6.3.1 FCN-based Segmentation 118

6.3.2 DeepMedic based Segmentation 122

6.3.3 Comparison to end-to-end Strategies 124

6.4 Discussion and Conclusion 126

7 Summary and Future Work **131**

7.1 Summary 131

7.2 Limitations and Future Work 134

Bibliography **135**

List of Tables

3.1 Predicting DSC for different segmentation methods using different RCA classifiers. Best results are obtained with the Single-Atlas approach yielding the lowest error between real and predicted DSC scores and high classification accuracy. . . 61

3.2 Detecting segmentation failure 66

3.3 Predicting different segmentation metrics 70

3.4 Comparison of predicting DSC with and without calibration via regression. . . . 70

4.1 Baseline and upper-bound accuracies. The upper-bound gives the highest performance when the network is trained and tested on one domain (target dataset). The performance drops significantly when training and testing data are from different domains (baseline). 80

4.2 Strategies of DARCA on liver segmentation: Training from scratch, fine-tuning, and iterative scheme (only for 2nd iteration with RCA selection) with different choice of subject selection. 83

4.3 Paired t-tests are performed for significance analysis against **Best-5 Worst-5** sample selection strategy by RCA, with the significance threshold at p= 0.05 . . 83

4.4 Baseline and upper-bound accuracies for right kidney segmentation 85

4.5 DARCA-FT on right kidney segmentation. Combination of best and worst subject selection gives the best result, and RCA selection also gives a similar accuracy to the real selection. 85

5.1 Multi-modal learning in MR images 100

5.2 Multi-modal learning in CT images 100

5.3 Overview on multi-modal learning on all organs 100

5.4 Swap stream of dual-stream Version 4 103

5.5 Random input for training dual-stream network 104

5.6 Multi-modal learning in MR images 105

5.7 Multi-modal learning in CT images 105

5.8 Overview on multi-modal learning on U-Net based on all organs 106

6.1 DSCs of coarse multi-organ segmentation by multi-atlas, baseline FCN , weighted FCN (W-FCN) and weighted-FCN with spatial prior (W-FCN-S). Small organs are shown in italics 120

6.2 Different strategies on segmentation of small organs, using the baseline FCN, weighted-FCN (W), auto-context (AC), two-stage networks (2SN), and combined multi-atlas (CMA) for ROI selection. Improvement in overall small organ segmentation accuracy (reported in DSCs) is achieved with our proposed approach. 120

6.3 Paired t-tests are performed for significance analysis against **FCN** as the baseline, with the significance threshold at p= 0.05 121

6.4 Paired t-tests are performed for significance analysis against **Two-stage FCN** as the baseline, with the significance threshold at p= 0.05 122

6.5 DSCs of coarse multi-organ segmentation by multi-atlas, baseline DeepMedic, weighted DM (W-DM) and weighted-DM with spatial prior (W-DM-S). Small organs are shown in italics 123

6.6 Different strategies on segmentation of small organs, using the baseline DM, weighted-DM (W), auto-context (AC), two-stage networks (2SN), and combined multi-atlas (CMA) for ROI selection. Improvement in overall small organ segmentation accuracy (reported in DSCs) is achieved with our proposed approach. 124

6.7 End-to-end strategies on segmentation of small organs. Organ segmentation Dice score of with DeepMedic (DM) as the baseline, compared to BASIC network, with dilated-convolutions (ASPPc and ASPPs) and autofocus layers (AFN) . . . 125

6.8 End-to-end strategies on segmentation of small organs. Organ segmentation Dice score of with DeepMedic weighted (DM-w) as the baseline, compared to BASIC network, with dilated-convolutions (ASPPc and ASPPs) and autofocus layers (AFN) 127

List of Figures

1.1 Visceral: internal organs of the human body, referring those within the chest, abdomen, and pelvis. Image taken from source http://nanoprom.info/ 4

1.2 Images from Dixon sequence: (A) In-phase. (B) Out-of-phase. (C) Fat-only. (D) Water-only. 7

1.3 Whole-body scans from MALIBO datasets, consists of T1w and T2w images. . . 9

2.1 Workflow for multi-atlas segmentation, image adapted from [119] 19

2.2 Box feature B_1 and B_2. Calculating the feature response of reference point $\mathbf{p}$ . . 23

2.3 Decision tree training 24

2.4 Decision tree testing 24

2.5 Workflow for random forest organ segmentation 25

2.6 Difference of the connection between convolutional and fully-connected layers. Right: Convolutional layer has sparse connection. Left: Fully-connected layer has dense connection, the image is adapted from [57]) 27

2.7 Difference of the the parameter sharing between convolutional and fully-connected layers. Top: Parameter sharing in convolutional layer. Bottom: fully-connected layer has no parameter sharing, so the parameter is used only once, the image is adapted from [57]) 28

2.8 The convolutional network is viewed as a small number of relatively complex layers, with each layer having many stages. (The figure is adapted from [57]) . . 29

2.9 Node graphs of 1D representations of CNN architectures (the figure is adapted from [90]) 29

2.10 Varying organ shapes in the population is shows in rendered 3D livers from MALIBO dataset 40

2.11 Automated quality control stages 41

2.12 Repeated station leads to wrong station content 41

2.13 Station Classification: Calculate MI between a query station and each of six-stations in reference. The highest MI score gives the right station name. 43

2.14 Inter-station (Left) and intra-station (Right) fat-water swap cases 43

2.15 Anomaly detection for each station 44

2.16 Overview of manifold learning technique: representing data in lower dimension by embedded manifold representation 46

3.1 Overview of the general framework for reverse classification accuracy. The accuracy of the predicted segmentation S_I of a new image I is estimated via a RCA classifier trained using S_I as pseudo ground truth, and applied to a set of reference images J_k for which ground truth reference segmentations $S_{J_k}^{\mathrm{GT}}$ are available. The best segmentation score computed on the reference images is used as prediction of the real accuracy of segmentation S_I. 58

3.2 Scatter plots of predicted and real DSC of multiple structures for three different segmentation methods (rows) using three different RCA classifiers (columns). High correlation and low prediction errors are obtained when employing the single-atlas label propagation as RCA classifier (right column). There is also good correlation with predictions in case of Atlas Forests (left) with larger spread towards lower quality segmentations. The constrained CNNs (middle column) are less suitable for RCA which is likely due to the difficulty of training on single images. Both, Atlas Forests and constrained CNNs work best for larger organs such as liver, lungs, and pelvis while leading to many zero predictions for smaller structures such as adrenal gland and clavicles. The single-atlas label propagation makes accurate predictions of segmentation quality across all 15 anatomical structures. A summary of the plots is given in Tab. 3.1. 62

3.3 Visual examples for predicting the segmentation quality of the liver for a new image shown on the top. Its GT segmentation (red colored) is unknown in practice, and we want to estimate the quality of the predicted, automatic segmentation shown on the most right (green colored). By taking the predicted segmentation as pseudo GT and training a RCA classifier we can obtain segmentations on a reference database with 24 images with available GT. The bar plot shows the real DSC in red and the different DSC values obtained for the reference images shown below. The green bar corresponds to the maximum DSC, and is selected as predicted DSC for the new image according to Eq. (3.1) which matches well the real DSC. 63

3.4 The bar plots show two examples for predicting the real DSC (red) in case of a good quality (left) and bad quality segmentation (right) using a database with 24 reference images with available GT. The predicted DSC (green bar) selected according to Eq. (3.1) matches well the real DSC. 64

3.5 Examples on predicting DSC from different segmentation methods on organ segmentation. RCA gives a good agreement to rank the best segmentation methods and to predict the DSCs in right lung segmentation (top) and liver segmentation (middle). Bottom: Prediction of low DSCs on pancreas segmentation also gives a good agreement to rank the best methods. 65

3.6 Scatter plots for the experiment for detecting segmentation failure when using degraded Random Forests with limited depth as the segmentation method. Atlas Forests (left) and single-atlas label propagation (right) make highly accurate predictions in the low DSC ranges and thus, are able to correctly detect such failed segmentations, with the exception of the bladder. Constrained CNNs are again less suitable for RCA with many zero predictions. 66

4.1 Examples of whole-body scans from source (MALIBO) and target (UK Biobank) database. 76

4.2 Overview of the strategies we tried for n-subject selection in DARCA: 1. Training from scratch. 2. Fine-tuning. 3. Iterative DARCA 78

4.3 Plot for different n-selection training size in different strategies: Fine-tuning (FT), fine-tuning with pseudo-labels (PL), and training from scratch (S+T). Similar trends are shown between real and RCA selection on FT and S+T with different size. 83

4.4 DARCA fine-tuning in liver segmentation. Combination of lowest and highest RCA prediction can give a better result than fine-tuning with random selection and with all of the target subjects. 84

4.5 DARCA fine-tuning in right kidney segmentation. Combination of lowest and highest RCA prediction can give a better result than fine-tuning with all of the target subjects. 85

5.1 Baseline network architecture for multi-organ segmentation: FCN based network with encoder-decoder structure 91

5.2 Left: Residual unit with pre-activation. Right: Upscore Unit. Input 1: Higher resolution features from the encoder (same scale) to add. Input 2: Input features from decoder to be upscored. 92

5.3 Dual-stream models. Version 1 - modality-specific encoder and shared decoder. Version 2 - shared part-of-encoder and decoder. Version 3 - shared encoder and modality-specific decoder. Version 4 - different streams in encoder, shared last layers of encoder, and modality-specific decoder. 94

5.4 Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow). 101

5.5 Multi-modal learning on CT. Liver (red), spleen (green), right kidney (blue), left kidney (yellow). 102

5.6 Boxplot of different organ segmentations on CT data: liver, spleen, right kidney, and left kidney segmentation (left-right). Different models on x-axis and Dice scores on y-axis. 102

5.7 Boxplot of different organ segmentations on MRI data: liver, spleen, right kidney, and left kidney segmentation (left-right). Different models on x-axis and Dice scores on y-axis. 103

5.8 Left: Residual unit with pre-activation. Right: Upsample and concat Unit. Input 1: Higher resolution features from the encoder (same scale) to add. Input 2: Input features from decoder to be upsampled and concatenated. 105

5.9 Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow). 106

5.10 Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow). 107

5.11 Boxplot of MR and CT in FCN based multimodal learning. 108

5.12 Boxplot of MR and CT in UNET based multimodal learning. 108

6.1 Overview. First stage: Coarse-scale segmentation with multi-organ segmentation with weighted-FCN, where we obtain the segmentation results and probability map for each organ. Second stage: Fine-scale binary segmentation per organ. The input consists of a cropped volume and a probability map from coarse segmentation. 115

6.2 Fine-scale segmentation of small organs (pancreas, gall-bladder and sacrum) with end-to-end AFN 126

6.3 Two-stage network (2SN) without weighting (W) and auto-context (AC) miss the local ROI selection on some organs (right clavicles and rectum), while adding spatial prior from multi-atlas (CMA) could enhance the segmentation results. . 128

6.4 Fine-scale segmentation of small organs (pancreas, recum, and bladder). Two-stage network (2SN) with weighting (W), auto-context (AC), and combined spatial prior from multi-atlas (CMA) can improve small organ segmentation results. 129

6.5 Fine-scale segmentation of small organs (pancreas, gall-bladder, rectum) in DeepMedic based segmentation. Two-stage network (2SN) with weighting (W), auto-context (AC), and combined spatial prior from multi-atlas (CMA) can improve small organ segmentation results. 129

6.6 Two-stage network (2SN) without weighting (W) and auto-context (AC) miss the local ROI selection with larger FOV on some organs (right clavicles and sacrum), while adding spatial prior from multi-atlas (CMA) could enhance the segmentation results. 130

Chapter 1

Introduction

1.1 Background and Motivation

In this section, we give a brief introduction to the anatomy of the human body. Later, whole-body imaging is discussed and we highlight the modalities and protocols that are used particularly in whole-body imaging. Similarly, we give a brief summary of MR whole-body imaging and its clinical role.

With the growing number of medical images produced, we need to adopt machine learning techniques for large-scale image analysis. One key challenge to extract useful clinical information from image datasets is the automatic understanding of image content which requires the separation of regions into semantically and anatomically meaningful parts. The task is also known as semantic image segmentation. In the clinical setting, trained radiologists read the scans to infer clinically useful information. But, this task is challenging when the number of scans is large and expert availability is limited.

In addition, manual image segmentation is a tedious and time-consuming task. There is a need for automated methods that can analyse medical images without requiring manual human interaction. The acceptance of automatic segmentation techniques in the clinical setting depends on the techniques being accurate, computationally efficient, and requiring minimal user inter-

action. With the advancement of machine learning, we hope to help this clinical work flow to be faster and more accurate.

One of the envisioned clinical applications of the thesis is to be able to automatically conduct the image segmentation quality control in unlabeled large scale medical image data. Furthermore, we also envision to have more accurate multi-organ segmentation accuracies in whole-body scans from different modalities, with the implementation of advance machine-learning strategies.

1.1.1 The Anatomy of Human Body

The human body consists of many individual structures, including bones and organs. Organs have a specific function within the body and reside within body cavities, such as the abdominal cavity (where the intestines, liver, and spleen sit) and pleural cavity (lungs and heart) [100]. Some organs work together to form different systems in our body, such as the digestive system, the respiratory system, and the urinary system. Anomalies or pathology found in one of the organs will disrupt the functionality of the body system. The musculoskeletal system, which gives the body basic structure and the ability of movement, consists of the human skeleton (bones, ligaments, tendons, and cartilage) and attached muscles. Abnormalities found on bones may affect the body movement and support.

With machine learning approaches, algorithms will be trained to recognise structures in whole-body scans, which include organs and bones. Most of the structures considered in our studies are included in the visceral anatomy, as shown in Figure 1.1. Descriptions of each structure included in this thesis work are detailed as the following [100]:

Organs:

- Heart: The heart pumps blood through the circulatory system supplying oxygen and nutrients to the cells of the body while removing carbon dioxide and other waste.
- Left and right lungs: In the respiratory system, the lungs help oxygen from the air we breathe enter the red cells in the blood, and help the body to get rid of CO_2 gas when

we breathe out.

- Liver: The liver detoxifies chemicals, metabolises drugs, and synthesizes proteins in the body that are required for many functions.

- Adrenal Glands: Endocrine glands that are located above the kidneys and produce a variety of hormones e.g. adrenaline, and the steroids aldosterone and cortisol.

- Kidneys: The kidneys help the body pass waste as urine, maintain overall fluid balance, and filter material from blood. The kidneys also secrete a number of hormones that help produce red blood cells, promote bone health, and regulate blood pressure.

- Gall bladder: A small organ that lies beneath the liver where bile is stored and concentrated before it is released into the small intestine.

- Spleen: An organ to the left of the stomach, which acts as a filter for blood as part of the immune system.

- Pancreas: Part of the digestive system, which produces hormones such as insulin (to help keep sugar levels in the blood stable) and digestive enzymes or "pancreatic juices".

- Bladder: An organ that stores urine, located in the pelvic area.

- Rectum : The final straight portion of the large intestine, which acts as a temporary storage site for fecal matter before it is eliminated from the body through the anal canal.

Bones:

- Pelvis: The lower part of the trunk of the human body between the abdomen and the thighs. The pelvic cavity is bounded by the bones of the pelvis and primarily contains reproductive organs.

- Spine: The spine provides the main support for the body (made of 33 individual, stacked bones) and protects the spinal cord from injury.

- Sacrum: Triangular bone at the base of the spine, that forms by the fusing of sacral vertebrae and articulates with the hip bones.
- Right and left clavicles: Located at the base of the neck to hold the arms freely and supported, so that the arm has maximum range of movement.

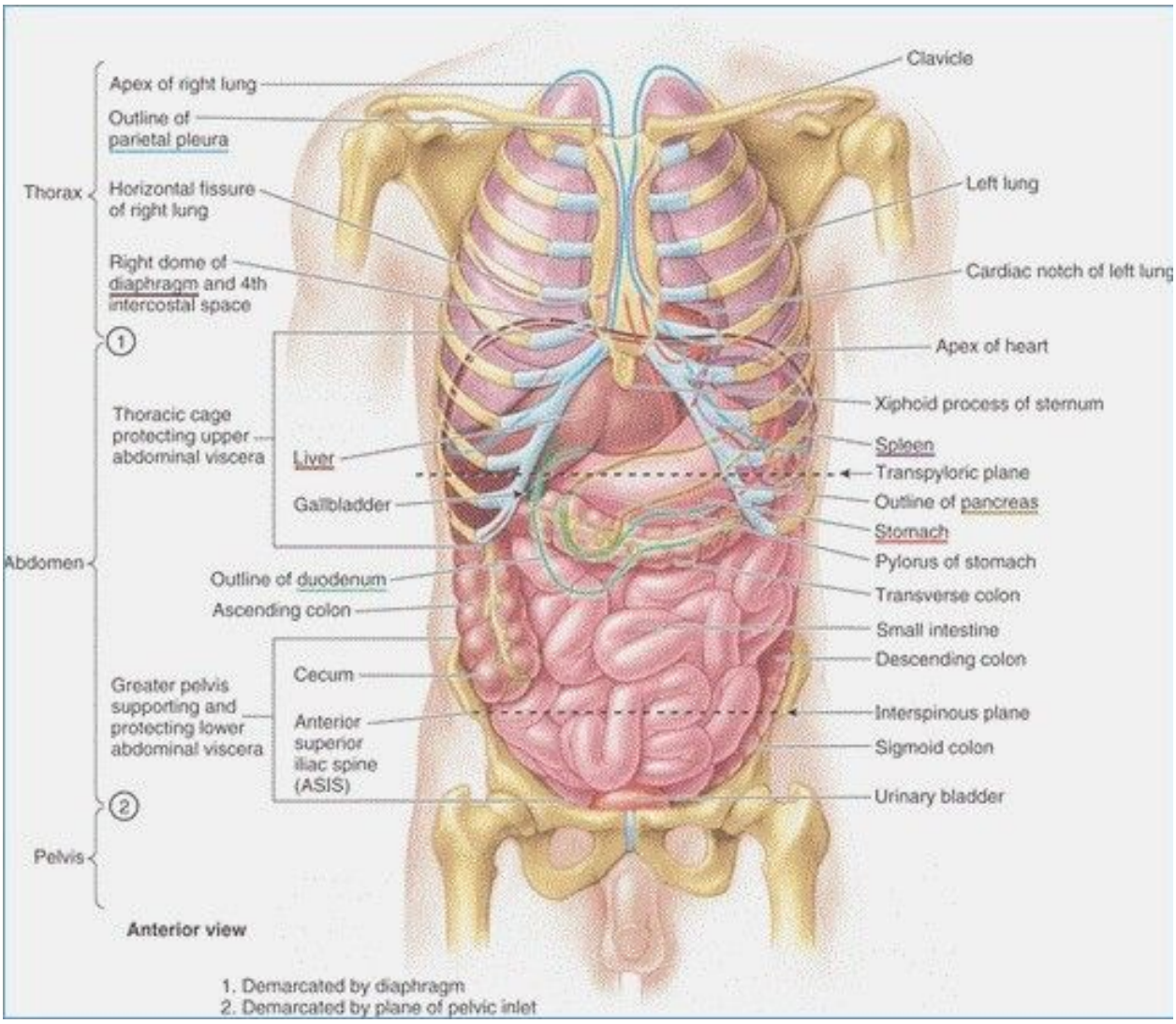


Figure 1.1: Visceral: internal organs of the human body, referring those within the chest, abdomen, and pelvis. Image taken from source http://nanoprom.info/

1.1.2 Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) is a non-invasive and non-ionising imaging technique that can produce tomographic images with an unmatched soft tissue contrast. The principle of

MRI is based on the absorption and emission of energy in Radio Frequency (RF) range of the electromagnetic spectrum [17]. MR images are produced based on spatial variations in the phase and frequency of the RF energy being absorbed and emitted by the anatomical tissues. Under the influence of a strong magnetic field, magnetically aligned hydrogen nuclei generate transverse magnetisation as a response to applied RF pulse sequences, which are captured by the scanner and then reconstructed as an image. The magnetic response of the tissues across time is shaped by atomic properties that differ depending on tissue type.

MRI makes use of strong magnetic fields, radio waves, and field gradients to generate detailed images of internal body structures. The water and fat molecules in the human body naturally consist of hydrogen atoms, which have their own magnetic fields and inherently spin around their axis at random orientations. An MRI scanner generates a strong magnetic field, B_0, that causes protons in hydrogen nuclei to align with the direction of the B_0 field. The nuclei precess around their axis and generate a longitudinal magnetisation. A RF pulse is directed at these precessing protons to temporarily create a second magnetic field perpendicular to B_0 [17].

Once the RF pulse is removed, the nuclei lose their magnetisation and reach an equilibrium, to recover their initial state with respect to the magnetic field B_0. Intensity values in these images result from the different concentration of hydrogen protons in different types of tissues. They are characterised by two relaxation factors: T1 (longitudinal) relaxation, which is the realignment of nuclei with the direction of the external magnetic field B_0, and T2 (transverse) relaxation, which is the loss of coherence among the nuclei and results in the decay of transverse magnetization. By changing the parameters of the MRI pulse sequence, different contrast images can be generated based on the T1 and T2 relaxation factors, e.g. T1-weighted (T1w) and T2-weighted (T2w) images, which can distinguish between different types of tissue [17].

MRI in Whole-body Imaging

MRI is chosen as imaging modality for whole-body scans because it is non-invasive, does not use radiation,not require a contrast agent or administration of medication, and has reasonable scanning time (less than 20 minutes) [111]. Whole-body MRI is the highest in dimensionality

and complexity since it covers almost all parts of a patient's body [121]. In general, MRI is a safer alternative to CT, as it has no ionizing radiation. Compared to CT, body MRI is a rich modality with multiple contrast mechanism (T1, T2 sequences, contrast-enhanced, etc). MRI also includes tools to characterize different type of tissues (water vs fat), such as Dixon technique, which we use in this study.

Whole-body MRI is recognized as an excellent imaging modality for imaging the solid organs (e.g. liver, pancreas, kidneys, adrenal glands and spleen) as well as the abdominal vasculature. MRI also provides a front-line diagnostic test for abdomen (liver, gall bladder, pancreas, kidneys, adrenal glands) and soft tissues. However, whole-body MRI is a multi-station acquisition, unlike CT that is capable of scanning entire chest, abdomen, and pelvis during a single breath-hold [121]. Thus, we have to use stitched images that combine images from multiple stations to form entire body images.

Whole-body MRI scans were obtained using Dixon technique [37]. This technique allows us to have four different contrasts of images (in-phase, out-of-phase, water only, and fat only) per subject [37]. Dixon technique uses phase differences to calculate water and fat components in post-processing. When fat and water are in-phase in the center of the echo, an in-phase image (IP) is acquired, vice-versa for creating an out-of-phase image (OP). From these in-phase and out-of-phase images, water only (fat-suppressed) image (W) and fat only image (F) can be reconstructed. Four different weighted-images of Dixon are shown in Figure 1.2 and the mathematical relation between four-sequences of Dixon is given below:

$$IP = W + F$$

$$OP = W - F$$

$$W = 0.5(IP + OP)$$

$$F = 0.5(IP - OP)$$

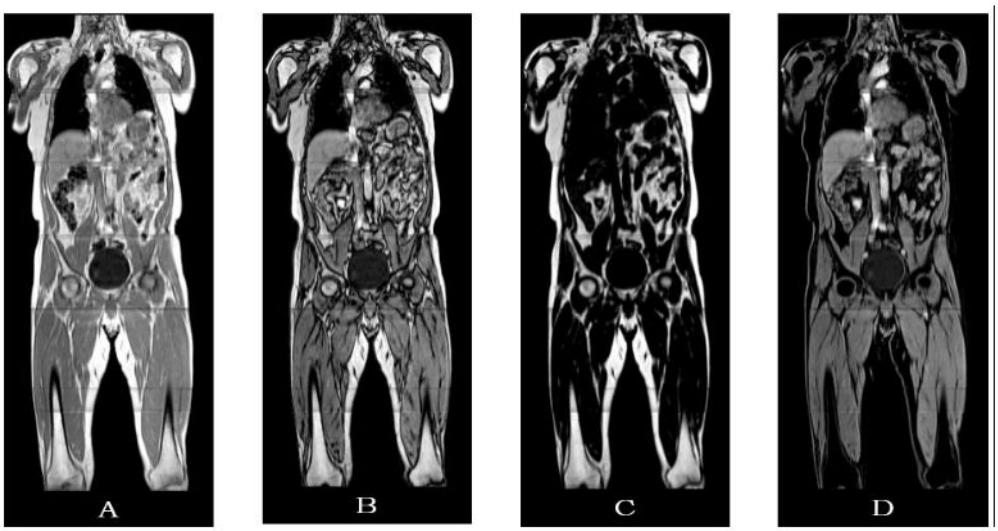


Figure 1.2: Images from Dixon sequence: (A) In-phase. (B) Out-of-phase. (C) Fat-only. (D) Water-only.

1.1.3 Computed Tomography (CT)

Computed Tomography (CT) imaging utilises X-rays to uncover internal body structures [122]. While lying on a table, a patient is moved slowly through a large hollow cylindrical X-ray machine, which contains a series of X-ray emitters and detectors. X-ray beams from different positions / angles traverse into patients body slices and are attenuated based on different absorption rates of the tissues. The combination of 2D X-ray images obtained from different positions and angles around the patient are processed to generate a 3D image via tomographic reconstruction.

This difference in absorption rates of the tissues make it possible for the X-ray to capture the different internal structures of the body. The intensity of a particular voxel of a CT image is proportional to the absorption rate of the X-ray beam that passed through that particular voxel location. Tissues in a CT image show up with varying intensities of gray. Intensities range from -1024 to +3071 HU (Hounsfield unit), which is a normalised measure relative to the absorption of water. The higher the value on the scale, the higher the absorption. Air-filled cavities (-1000 HU) appear as black on the CT image, while bone structures (+1000 HU) appear as white. If the CT imaging machines are properly calibrated, then there should be little to no difference

in intensities between machines [122].

CT is an important imaging tool and is heavily used for the diagnosis of disease because it is fast to acquire and can provide high resolution images. CT imaging is less prone to motion artefacts due to the fast acquisition speed; however, it can still suffer from other artefacts, such as noise. The radiation exposure time and dose may also have an influence on the image quality. Moreover, X-rays are a type of ionising radiation and can be harmful when a person is exposed to a high dose.

1.2 Datasets Used

In this work, we predominantly use the MRI whole-body scans from MAchine Learning In whole Body Oncology (MALIBO) datasets [114]. However, in some of the experiments, we also use MRI images from the UK Biobank datasets [130] and CT abdominal datasets, which will be described as follows:

1.2.1 MRI Whole-body Images from MALIBO dataset

Our in-house MALIBO datasets consists of whole-body, multi-sequence MRI (T1w Dixon and T2w images) of 35 healthy volunteers. Whole body MRI was performed in a moving-table 1.5 T system (Siemens Avanto with Syngo MR B17, Erlangen, Germany), using the body coil for transmission and the neck/ body phased array coils as receive coils. Manual annotations of 15 anatomical structures, including abdominal organs (heart, left/right lung, liver, adrenal gland, gall bladder, left/right kidney, spleen, pancreas, bladder) and bones (spine, left/right clavicle, pelvis) have been generated by clinical experts as part of the study. These manual annotations will serve as ground truth (GT) in the quantitative evaluation.

As the datasets keep growing, for the latter experiment we use MALIBO-HV data. This dataset consists of 48 healthy subjects with 18 anatomical structures annotated, including the previous 15 with added rectum, sacrum, and humerus. Complete scans from one patient in

MALIBO data consist of T1w (fat, water, in-phase, out-of-phase) and T2w. Figure 1.3 shows that T1w and T2w images from a patient. However, for a more efficient and better intensity interpretation, we use only T2w images from MALIBO dataset in our experiments, as the manual annotation were generated on T2w images.Hence, alignment error between images and labels could be reduced.

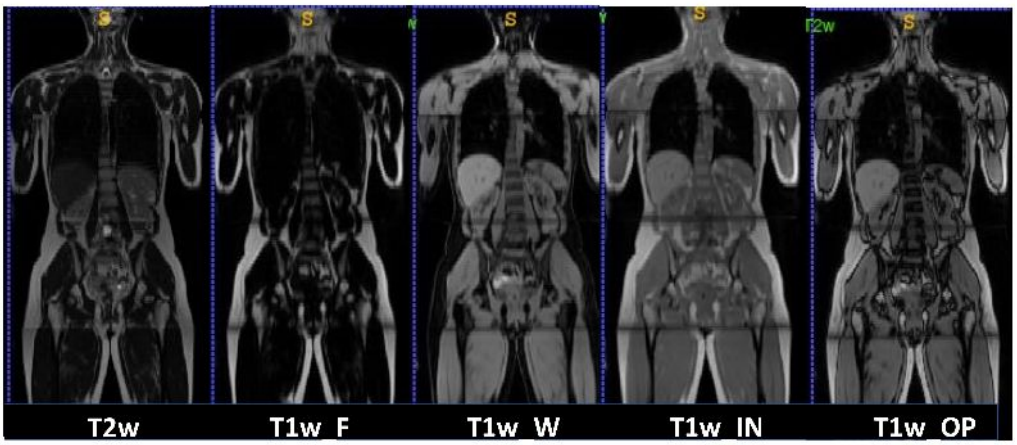


Figure 1.3: Whole-body scans from MALIBO datasets, consists of T1w and T2w images.

1.2.2 MRI Whole-body Images from UK Biobank dataset

With the initiative from the UK Biobank [130], large-scale medical data are now digitally available. UK Biobank is a large prospective study with over 500,000 participants of middle and old age (40-69 years). Between 2006 - 2010, the study has collected extensive medical data from participants, such as questionnaires, physical measures, habitual data, and biological samples for a wide range of health-related outcomes. Particularly, UK Biobank Imaging study that started in 2014 is targeting to scan 100,000 subject from the UK population over a 5-6 year period. At the moment, around 32,000 participants were successfully scanned. UK Biobank Imaging study gathers comprehensive imaging phenotypes, including brain, cardiac, whole-body MRI (Magnetic Resonance Imaging), 3D carotid ultrasound, and DEXA from participants. The integration of large-scale data in UK Biobank (from body-part imaging and non-imaging data)

will support the investigation of diseases, leading to improvement in prevention, diagnosis, and treatment [111].

Whole-body scans in the UK Biobank were acquired from 1.5 T MRI, with T1-weighted with 10 mm slice thickness, 10 mm gaps between slices [130]. Dixon whole-body sequence generates four weighted images of T1w (F, W, IN, OPP) for each subject, as shown in Figure 1.2. Originally, data were acquired in the form of six stations for each subject. Then, stitching images were generated by physicians. Whole body (stitched) image corresponds to six stations, covering from shoulders to knees. All of the UK Biobank scans are unlabeled, without manual annotations. For our experiments, we only use the first acquired 100 subjects of whole-body scans from UK Biobank dataset.

1.2.3 CT Abdominal Images

We use the publicly available dataset of MICCAI 2015 Workshop and Challenge on Multi-Atlas Labeling Beyond the Cranial Vault (`https://www.synapse.org/#!Synapse:syn3193805/wiki/218441`). The dataset contains abdominal anatomy and cervix, from 30 patients with manual annotation for 13 organs. The 3D scans exhibit typical contrast variations between organs. CT scans in the dataset have in-plane resolution between (152, 152, 210) to (175, 175, 245) in size, the number of slices is between 238 and 1061, and 2 mm isotropic spacing.

1.3 Technical Challenges in Whole-body Image Analysis

Machine learning has been applied in various application on medical image analysis. However, there are several challenges, particularly for applying machine learning on whole-body scans, such as high variability of different structures in the body and small number of annotated datasets, compared to other datasets. The low contrast between different structures and soft tissues in whole-body scans results in a difficult detection task.

<u>High variability of anatomical appearance</u>

Whole-body MRI is not only high-dimensional (covers almost all parts of human body) but also complex. Compared to single organ scans (heart or brain scans), whole-body scans have high variability. Whole-body contains multiple structures with different size, shapes, and intensity that makes them challenging to be segmented correctly. Across a population, there is also a wide variety of shapes and appearances of organs between subjects. In addition, some organs are more difficult to segment than others, e.g. the pancreas is difficult to segment due to its small size and deformable nature, while the liver, on the other hand, is easier to segment due to its large size.

Image intensity in whole-body images

Some internal organs may have similar intensities in whole-body MRI. In addition, the range of intensities of a given organ can vary from patient to patient. Thus, it is difficult to simply rely on intensity models in segmenting an organ. The difficulty is due to the inhomogeneous intensity variations within MRI scans. Unlike CT, MRI images do not have standard relationship between tissue type and image intensity (compared to Hounsfield units in CT). Internal organs being in close proximity to each other also sometimes have poor edge contrast with weak or non-existent boundary between them, for example the pancreas and the adrenal glands. Therefore, it is very hard to distinguish that particular organ from its surrounding.

Unavailability of large-scale annotated data

Although, we have large-scale datasets, such as the UK Biobank, they are all unlabeled. Manually annotating one whole-body image is already tedious as it has more than 15 important structures of interest. The unavailability of ground truth brings the challenge for evaluating segmentation accuracy from automated methods. For MALIBO datasets, even though the acquisition was done at the same local imaging center (Hammersmith Hospital Campus), still some corrupted data may appear. Corrupted data, wrong annotation, swapped stations, and low-quality images with artefacts that cannot be used for further analysis reduced our number of samples.

Different imaging modalities used

In some hospitals, MRI scanners might be available, while in other hospitals, CT is more

frequently used. Here, there is a need to learn effectively from multi-modalities. Thus, with more data, even though from unpaired multi-modal images, the automated segmentation results could be made more reliable.

Variation between imaging centers

Different hospitals or scanners used may result in different image appearance. Segmenting scans acquired from different imaging centers than the one use for algorithm training, can lead to degraded segmentation results. There is a need to address a more efficient transfer learning strategy between labeled database to a larger unlabeled imaging database.

Computational requirements

Current deep-learning based techniques requires expensive computation and longer processing time. Additionally, medical images are in 3D, which can be an obstacle when dealing with a large dataset or bigger network, when computational resources are scarce, or when close to real-time results are required.

1.4 Thesis Contributions

The following section presents the contributions of this thesis in analysis of whole-body imaging data and also summarises how the aforementioned challenges are addressed.

Predicting Segmentation Accuracy without Ground-truth

We propose a novel technique Reverse Classification Accuracy (RCA) to address the segmentation evaluation in the absence of ground-truth. Our method applied a single-atlas registration strategy as a reverse classifier to predict the accuracy score of an image segmentation on a per case basis. We demonstrate its applicability for predicting multi-organ segmentation on whole-body MRI scans.

Domain Adaptation using RCA

One of the important clinical applications of RCA is on domain adaptation. Difference in image appearances between different imaging centers brings the necessity of domain adaptation,

in order to get a similar segmentation quality across domains. We select the most useful samples from unlabeled target domain using RCA to be incorporated into the supervised domain adaptation scheme. Our proposed framework shows consistent results between different organs and different domains.

Multi-modal Learning on Unpaired Images

In practice, different hospitals may use different imaging modalities to scan the whole-body, such as MRI and CT. The images acquired from those modalities are distinct (in terms of intensity and appearance), but may preserve the basic features for the same organs. In that regard, we propose a novel approach to learn from unpaired images for MRI and CT. The presented multi-modal learning framework can learn the shared as well as the modality-independent features. We show that with our proposed dualstream-network, the segmentation of multi-organs can be improved in both modalities.

Small organ segmentation

The current state-of-the-art medical image segmentation is based on Convolutional Neural Networks (CNN) models. However, commonly used CNN architectures still show low accuracy when segmenting small organs in the whole body. This could be a problem as these organs, such as pancreas, bladder, etc. are not less important in diagnosis. A coarse-to-fine strategy with weighting schemes is proposed to incorporate spatial prior and visual cues into a standard CNN model. With this approach, the fine-scale network is guided to focus on segmenting a particular small organ in a local Region of Interest (ROI). Our solution demonstrates improved accuracy on small organ segmentation.

1.5 Thesis Overview

In the following chapter, background information regarding the whole-body image analysis problems and state-of-the-art methods are reviewed with a particular focus on machine learning. In particular we described the basics of segmentation algorithms that will be used in the proposed methods. We also discuss some of the more traditional machine learning applications

in population image analysis with whole-body scans.

Chapter 3 introduces RCA for predicting the segmentation accuracy for unlabeled data. Results are presented for multi-organ segmentation problems with 15 structures of interest, and compared against other reverse classifiers. In Chapter 4, RCA is extended to an application in supervised domain adaptation. Its applicability in selecting useful samples in unlabeled target domain is demonstrated with the experiment using different datasets acquired from different imaging centers. The idea of multi-modal learning is explored in Chapter 5. Its benefits for learning to segment multiple organs from unpaired images is shown in both modalities (CT and MRI). In Chapter 6, a coarse-to-fine strategy is proposed to segment small organs with weighting schemes, spatial prior and visual cues, which encourages CNN models to choose a more accurate ROI. Finally, in the last chapter, we summarise the work presented in this thesis and discuss future work.

1.6 List of Publications

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*, 36(8):1597-1606, 2017.

R. Robinson, **Vanya V. Valindria**, W. Bai, H. Suzuki, P.M. Matthews, C. Page, D. Rueckert, Ben Glocker. Automatic Quality Control of Cardiac MRI Segmentation in Large-Scale Population Imaging. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 720-727, 2017.

Vanya V. Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker. Multi-Modal Learning from Unpaired Images: Application to Multi-Organ Segmentation in CT and MRI. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 547-556. 2018.

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker. Domain Adaptation for MRI Organ Segmentation using Reverse Classification Accuracy. *1st International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

Robert Robinson, O.Oktay, W. Bai, **Vanya V. Valindria**, M. M. Sanghvi, N. Aung, J.M.Paiva, F. Zemrak, K.Fung, E.Lukaschuk, A.M. Lee, V.Carapella, Y.J.Kim, B.Kainz, S.K.Piechnik, S.Neubauer, S.E. Petersen, C.Page, D. Rueckert, Ben Glocker. Real-Time Prediction of Segmentation Quality. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 578-585, 2018.

Vanya V. Valindria, Ioannis Lavdas, Juan Cerrolaza, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker. Small Organ Segmentation in Whole-body MRI using a Two-stage FCN and Weighting Schemes. *9th International Conference on Machine Learning in Medical Imaging (MLMI)*, pages 346-354, 2018.

Robinson, R.,**Vanya V. Valindria**, Wenjia Bai, Ozan Oktay, Bernhard Kainz, Hideaki Suzuki, Mihir M. Sanghvi, Nay Aung, Jos Miguel Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Chris Page, Paul M. Matthews, Daniel Rueckert and Ben Glocker. Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. *Journal of Cardiovascular Magnetic Resonance*, 21: 18, 2019.

Chapter 2

Review of Machine Learning Methods in Whole-Body Image Analysis

Computer-aided image analysis frameworks have been in development for many years to provide better clinical diagnosis from medical images. Machine learning approaches map input data into an abstract representation, in order to exploit the connection between input and output. An example of this is malignant or benign tissue classification. Recently, development of machine learning techniques has exponentially increased, with increasing interest from the medical imaging community. Computer-aided medical image analysis has started to adopt machine learning techniques for a faster and more accurate outcomes, such as lesion segmentation or tissue classification.

Traditional image analysis techniques rely on hand-crafted feature extraction that are then used as inputs to algorithms. However, these methods are outperformed by recent machine learning based approaches (particularly, deep learning) in several image classification benchmarks (e.g. PASCAL VOC 2012) . With the availability of large-scale medical imaging data and annotations, image analysis techniques have shifted towards learning-based methods.

Researchers have used machine learning approaches for various tasks, such as segmentation [74], landmark detection [2], survival prediction [28], image registration [66], pathology detection [123], image super resolution [104], to image synthesis [102]. In the context of whole-body

imaging, the most dominant application is still image segmentation [90], regardless of whether it is for organ segmentation or lesion segmentation.

The main factors that determine the performance of these machine learning approaches include; the availability of training data, the representation power of the neural network architecture, and the generalisation ability. In the last few years, whole-body imaging datasets have become available to support learning-based approaches, such as UK Biobank Imaging Study [130] and Visceral [69].

This chapter gives a brief overview of machine learning techniques and their application in whole-body imaging for various tasks, such as; segmentation, transfer learning, quality control, and population study. These methods share common goals: objective quantitative assessment, time-efficiency, and information extraction from imaging data. We first present commonly used methods in whole-body image segmentation and discuss their aspects in terms of model parameterisation, architecture design, and computational requirements. An overview of machine learning applications on population based whole-body imaging is also provided, which include image quality control.

2.1 Whole-Body Image Segmentation

Organ segmentation is a relatively well-studied topic, as many methods have been proposed to tackle this task. Through computer-aided image analysis, multiple organs can be objectively measured and evaluated for reliable quantification. There are multiple algorithms that have shown promising results in segmentation of different anatomical structures.

In terms of multi-organ segmentation, a prior challenge VISCERAL (Visual Concept Extraction Challenge in Radiology) conducted an extensive challenge on CT and MR whole-body cloud-based datasets [69]. It is dominated by more traditional methods, such as shape-appearance models and multi-atlas registration [69]. Their benchmarks showed that segmenting organs in MRI data is much harder than in CT, with only few organs that could be reliably detected. Previous works on multi-organ segmentation have applied more traditional approaches from

the computer vision literature, such as rich hand-crafted feature descriptors and supervoxels [172], context integration [85] and space learning [171, 160]. However, these approaches still struggle to segment organs with high variation and complex contextual information.

We will review the three popular multi-organ segmentation methods below, ordered based on their performace.

2.1.1 Multi-atlas Segmentation

Atlas-based approaches are among the most popular methods for multi-organ segmentation. Atlas-based methods typically require the use of image registration to align the atlas image to an unseen image. The notion of single atlas has been extended to multi-atlas segmentation, which employs multiple atlases to better capture anatomical variation. This has shown to improve single-atlas based segmentation accuracy significantly. The set of atlases can model anatomical variability, which make it more robust for image segmentation across a diverse set of images. Each atlas is registered to the target image, with the propagated labels from multiple atlases combined by label fusion. The consensus segmentation produced by multiple atlases is less likely to be influenced by registration errors, compared to registration using an individual atlas. An overview of multi-atlas segmentation method is depicted in Figure 2.1.

Two crucial steps in multi-atlas segmentation are; the selection of similar atlases, and accurate pairwise registration (between atlases and target image). This can be challenging in the case where there is high inter-subject variability. Furthermore, the use of multiple atlases requires high computation time. Thus, multi-atlas segmentation can be very expensive in terms of computing time. Furthermore, all atlases must be available during segmentation time.

The combination of multi-atlas registration and hierarchical weighting schemes (global, organ, and voxel level) can segment multiple organs reliably [149, 7]. An improved accuracy of atlas-based multi-organ segmentation can be achieved by incorporating discriminative dictionary learning for segmentation (DDLS) with voxel-wise local atlas selection strategy, and graph-cuts method on post-processing [134].

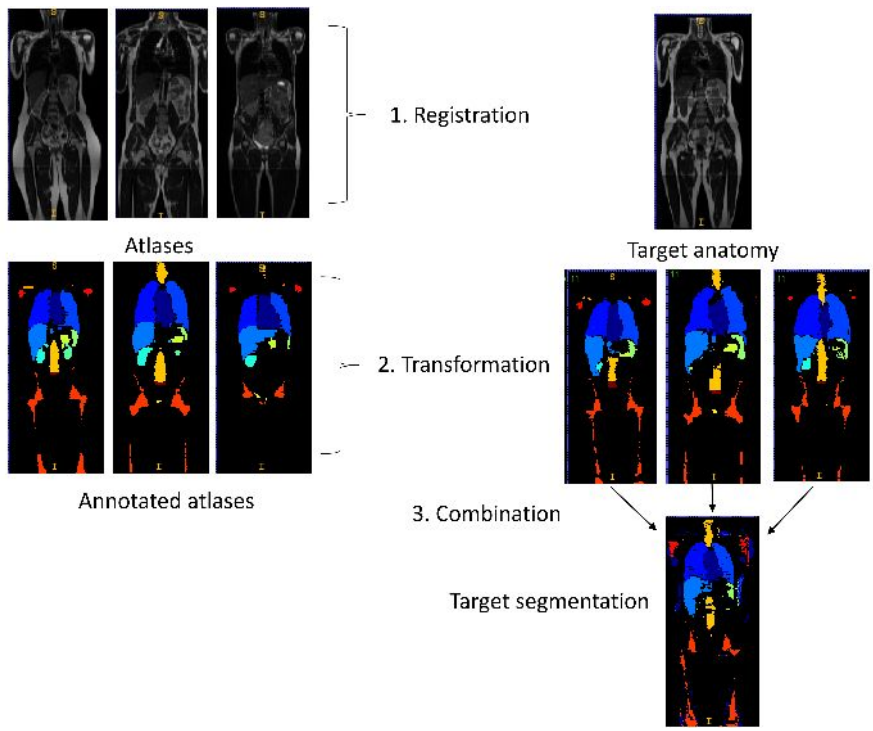


Figure 2.1: Workflow for multi-atlas segmentation, image adapted from [119]

Different label fusion strategies can be employed to combine the propagated labels from multiple atlas into a consensus segmentation. Several label fusion strategies from atlases were studied [68]. The first strategy is label fusion with majority voting (MV), which simply uses the maximum votes per voxel. Second, is weighted voting label fusion (LW), where the similarity of each atlas contributes to the final result with a weight. Third, patch-based label fusion (PB). Since all the atlases are at least roughly registered to the target image, we only consider for similar patches in a small neighborhood, instead of the whole image. Patch-based segmentation is essentially a weighted K nearest neighbours (KNN) classifier; if the atlas patch and the target patch look similar, then they should have a similar label. This technique evaluates the intensity similarity between the image patch centered at the target voxel and the atlas voxel to determine the weight. Fourth, patch-based label-fusion with augmented features (PBAF), which incorporates the augmented feature vector into patch-based segmentation in order to describe two patches more discriminatively. Fifth, label fusion with augmented feature vector using the support vector machine classifier (SVM-AF), instead of the KNN classifier in conventional PB label fusion.

However, multi-atlas segmentation has several drawbacks; such as lack of atlases, difficulty

for optimal atlas selection, and high cost of atlas-to-image registration. Hence, random forest methods are applied for multi-organ segmentation to overcome these drawbacks using context-rich visual features [32]. Another solution is by atlas forests, which was proposed in [169]. Atlas forest is random-forest based approach which encodes individual atlases for multi-atlas label propagation. This method performs well in both accuracy and low computational time as it only needs one registration per target image.

2.1.2 Random Forest

Ensemble methods became popular with the face detection application of Viola-Jones [145]. Random forest is an ensemble classifier, which is constructed by combining many decision trees. Each tree casts a vote for the most likely prediction, e.g., a segmentation class label for each image point [16]. The ensemble of different trees yields a more robust prediction with better generalization that produces a higher accuracy. Compared to another ensemble-based classifier, like Adaboost [43], random forest selects features randomly to determine split nodes of the decision trees, greedily searching for the best split with lowest error rates. Unlike random forest, Adaboost has no random elements but only the successive re-weightings of training sets. Adaboost algorithm builds a strong classifier by combining weak classifiers with linear weights.

The popularity of random forests for use in computer vision and medical imaging is reflected by its numerous successful applications, ranging from; pedestrian tracking, semantic segmentation, automatic parsing of medical scans, to human pose tracking [31]. Models of random forest have evolved (with other forest variants, such as random ferns and extreme random trees) and can be applied to various tasks, from manifold forest, classification forest, to regression forest [33],

Similar to other trainable classifiers, random forests involve a training phase (off-line) and a testing phase (on-line). By the end of the training phase, we expect to get optimum weak learners at each node, a learned tree structure, and different subsets of training points at each leaf. During testing, test function splits nodes until the data point reaches a leaf node which contain predictions (labels).

Training. The process of training a decision tree (off-line) in a random decision forest is described in Figure 2.3. The training set for the root node in a tree is S_0, which consists of input data v(collection of features) and their class/category c. Features are then split in split node S_j. Each node has parameters $\theta = (\phi, \psi, \tau)$, where $\phi(\mathrm{v})$ is the feature selection function, ψ is the decision parameter, and τ is the decision boundary. Selecting the type and parameters of the test function $h(\mathrm{v}, \theta_j)$ is important during the training phase. This function $h(\mathrm{v}, \theta_j)$ is also known as the weak learner model to split the data in node j.

At each node j, we learn the function that best split S_j to S_j^L and S_j^R, to left and right child nodes respectively, by maximizing an objective function. By training objective functions, we reduce uncertainty using the concepts of entropy and information gain.

The objective function at node j is:

$$\theta_j = arg\,max\, I_j \tag{2.1}$$

where

$$I_j = H(S) - \sum_{i \in \{L,R\}} \frac{S_j^i}{S_j} H(S_j) \tag{2.2}$$

I_j captures some notion of information gain from the split, as the Shannon cross-entropy is defined by:

$$H(S) = -\sum_{c \in C} p(c) log(p(c)) \tag{2.3}$$

where S is the set of training points, c indicates the class label, C denotes the set of all classes, and $p(c)$ is the empirical distribution extracted from the training points in S. Maximizing information gain gives us split parameters with the highest confidence in the final distributions.

Randomness is injected in the training phase of the random forest to give a better generalisation.

Two common approaches to do this are by:

- Bagging

  Randomly sub-sample training data for good training efficiency.

- Randomize Node Optimization (RNO)

  Denote that T is all possible parameter settings $\theta = (\phi, \psi, \tau)$. The idea is at node j, use a small random subset of $T_j \subset T$ of parameter values. Then, define the 'randomness' $\rho = |T_j|$, which controls the correlation between trees in the forest (where $\rho = 1$ means high randomness and $\rho = |T|$ returns identical trees). RNO injects randomness in training trees, as the child nodes receive a different subset of training set.

The tree grows branches until it reaches the stopping criteria, either the maximum tree depth (D) is reached or there are too few training samples available at a node [33].

An application of random forest in organ segmentation is effectively approached by utilising box features $\mathbf{B}$ [32], which capture the features and appearance of anatomical structures as well as spatial context. We wish to estimate the position of the 3D bounding box containing the organ of interest. The 3D feature box $\mathbf{B}$ is displaced from the reference point $\mathbf{p}$, where we can have a number of possible feature boxes. Each voxel $\mathbf{p}$ in an image votes for where it thinks the organ should be via relative displacement. The feature response $\mathbf{v(p)} = (v_1, ..., v_i, ..., v_d)$ for a reference point $\mathbf{p}$ can be calculated as the sum of mean intensity values over displaced feature boxes $\mathbf{B}_i$, which is defined by:

$$v_i = \frac{1}{|\mathbf{B_i}|} \sum_{q \in \mathbf{B_i}} J(\mathbf{q}) \tag{2.4}$$

In Figure 2.2, the box $\mathbf{B_p}$ is displaced relative to the reference point $\mathbf{p}$ by a given offset. $J(\mathbf{q})$ defines the image intensity at position $\mathbf{q}$. For each reference point, thousands of box features are randomly generated during training stage [33]. The randomness is injected during training, by giving subsets of random features available at each node, to improve the generalization.

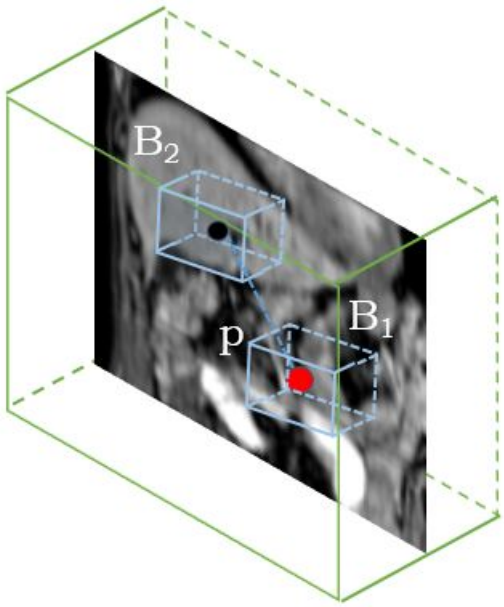


Figure 2.2: Box feature B_1 and B_2. Calculating the feature response of reference point $\mathbf{p}$

The datasets are manually labeled with 3D boxes around the organs of interest. A forest is trained to learn the association of voxel features with respect to bounding box location. Training stops when it reaches the stopping criteria, either at maximum depth level or the leaf only have minimum number of samples (specified before).

Testing. After the training stage, each leaf is left with a subset of (labeled) training data. During training, the leaf predictor model $p(c|\mathrm{v})$ is also learned, where v is the data point tested in the tree and c are labels. This posterior distribution $p(c|\mathrm{v})$ captured the leaf statistic, which can be useful to generate a label for that point. During testing, features are split based on test function $h(\mathrm{v}, \theta_j)$. This makes the input test data likely to end up in a leaf containing similar training points, which have similar associated labels. The process of testing (on-line) stage is shown in Figure 2.4 .

A similar process is applied during the training phase for multi-organ segmentation using random forest. The extracted features are pushed to trees and the split function applied. Then, it stops when it reaches its leaf node. The probability maps show the confidence level of classified features as particular labels. Label maps output gives the classification of features, as shown in Figure 2.5.

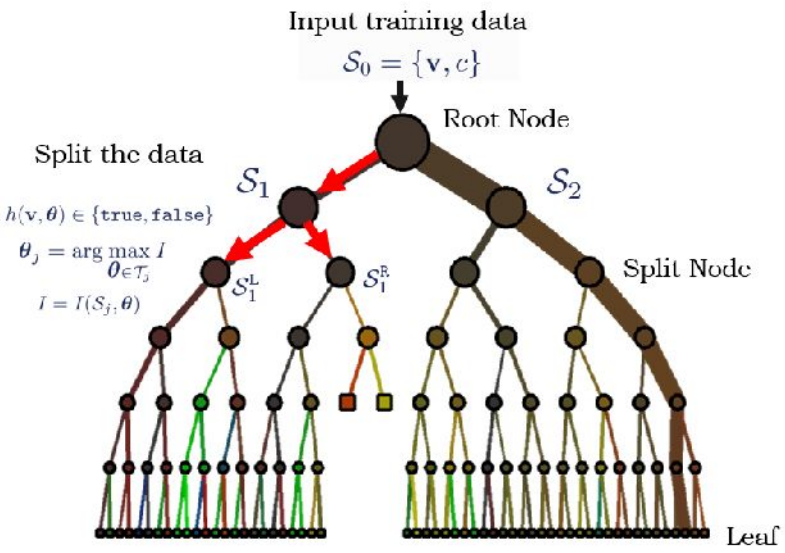


Figure 2.3: Decision tree training

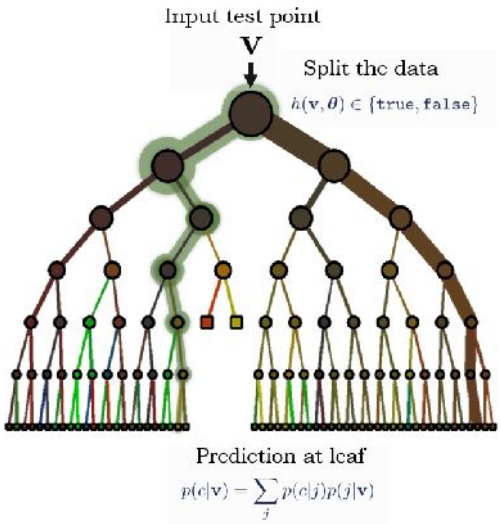


Figure 2.4: Decision tree testing

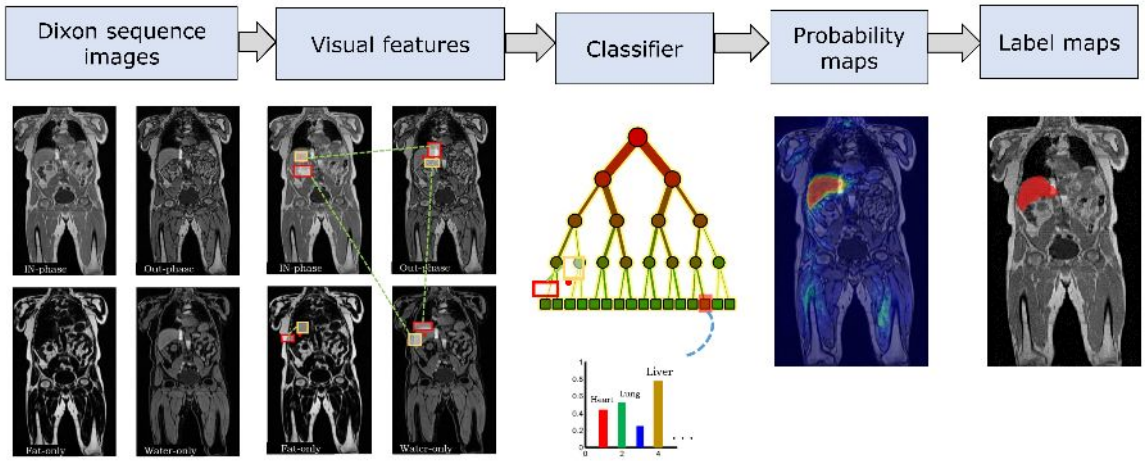


Figure 2.5: Workflow for random forest organ segmentation

There are some key parameters that can affect the behaviour of random decision forest [33].:

- Stopping criteria/tree depth (D), very deep trees can lead to overfitting.
- Tree randomness (ρ), which affects tree correlation and generalization properties
- Forest size (number of trees) T
- Choice of weak learner mode (h)
- Training objective function (I)
- Choice of features in practical applications (for example; the use of box features for organ localisation)

Prior to multi-organ segmentation, [30] predicted the location of the organ bounding box using regression forests. Their results show that regression forest can localize major anatomical structures with higher accuracy, robustness, and speed than atlas-based method. Speed is an important advantage of random forest-based methods. The work by Gauriau et al. [47] takes into account the spatial relationship and shape priors in the implementation of cascade forest regression for a refined multi-organ localization with relatively fast computational time.

2.1.3 Convolutional Neural Networks (CNN)

A convolutional neural network (CNN) is a feed-forward artificial neural network model that has been successfully applied in a number image analysis applications including classification, detection and segmentation. Convolution leverages three properties that improve a machine learning system [57]:

- Sparse interactions: The difference of connectivity patterns between convolutional and fully-connected layer is shown in Figure 2.6. Fully-connected layer has dense connections, which connects all inputs. In contrast, convolutional layer has sparse connections because only the local neighbors of pixels are connected, instead of using all pixels as input. Sparse connection is accomplished by making the kernel smaller than the input.

  As shown in Figure 2.6, if we view from below, we highlight input x_3 and see which output unit s is affected. In convolutional layer, as s is formed by convolution with a kernel of width 3, only three output are affected by x_3. Meanwhile, in fully-connected layer, all outputs s are affected by x_3 because of the dense connection.

  If we view from above the connections in Figure 2.6, we highlight one output s_3 and inputs x that affect this unit - these units are known as the receptive field of s_3. When s_3 is formed by convolution with a kernel of width 3, only three inputs affect s_3. As the connection in fully-connected layer is no longer sparse, all inputs affects s_3.

  Hence, sparse connectivity in convolutional layer reduces the number of parameters stored, which both reduces the memory requirements of the model and improves the efficiency.

- Parameter sharing: Parameter sharing refers to using the same parameter for more than one function in a model. In a convolutional layer, parameter sharing by convolution operation means that rather than learning a separate set of parameters for every location, we learn only one set. Convolution is more efficient than dense matrix multiplication in fully-connected layer in terms of the memory requirements and statistical eciency.

  Figure 2.7 shows how parameter sharing works. Black arrows in the figure are the connections that use a particular parameter. In a convolutional layer, the black arrows indicate

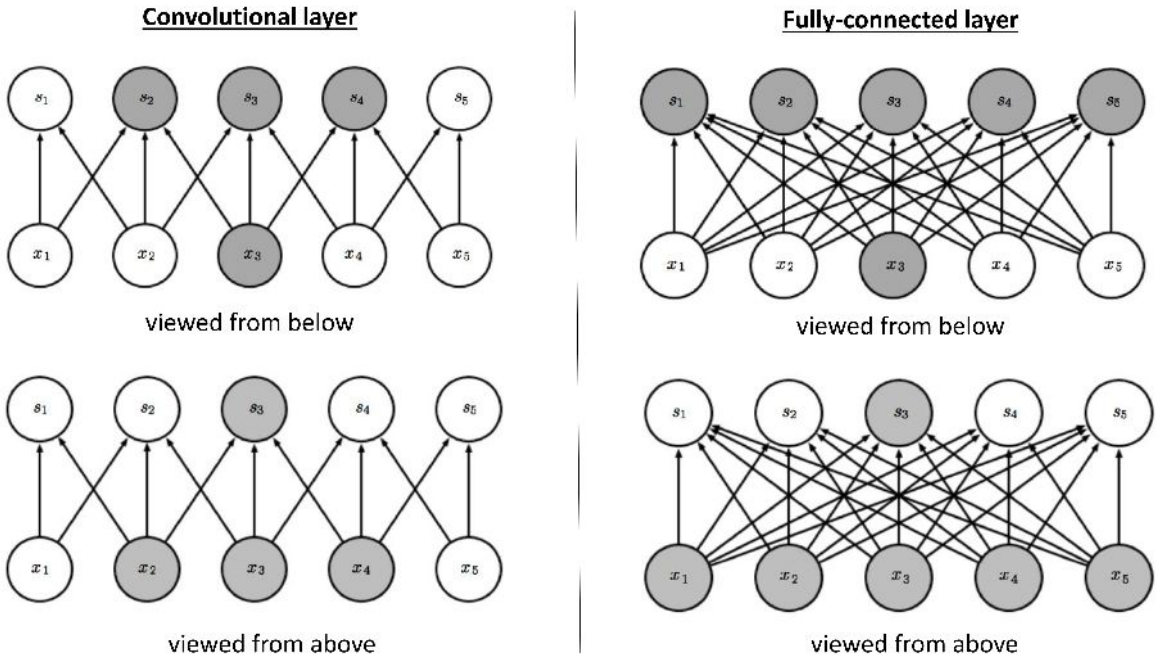


Figure 2.6: Difference of the connection between convolutional and fully-connected layers. Right: Convolutional layer has sparse connection. Left: Fully-connected layer has dense connection, the image is adapted from [57])

the uses of the central element in a kernel of width-3. This single parameter is used at all input locations because of parameter sharing. Meanwhile, in fully-connected layer, the single black arrow indicates the use of the central element of the weight matrix. This model has no parameter sharing, so the parameter is used only once and never revisited.

- Equivariant: it means if the input changes, the output changes in the same way.

A typical layer of a convolutional network consists of three stages (see Figure 2.8):

1. Convolution stage: Convolution layer makes use of a set of learnable filters to detect the presence of specific features present in the original input. This filter is convolved across the width and height of the input file, and a dot product is computed to give an activation map.

2. Activation stage: The activation function is the non linear transformation over the input signal.

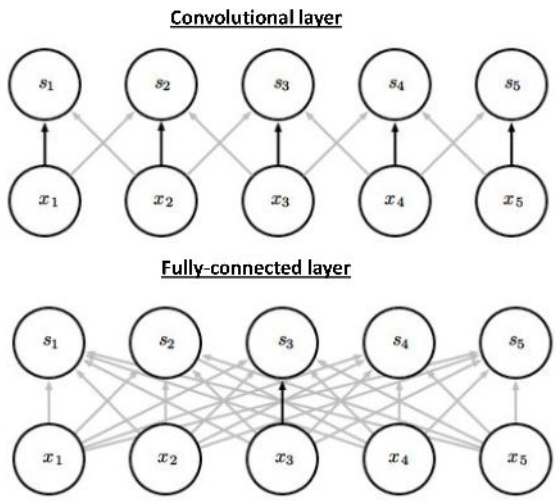


Figure 2.7: Difference of the the parameter sharing between convolutional and fully-connected layers. Top: Parameter sharing in convolutional layer. Bottom: fully-connected layer has no parameter sharing, so the parameter is used only once, the image is adapted from [57])

3. Pooling stage: Pooling function is used to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network. Pooling layer operates on each feature map independently.

A detailed example of a 1D CNN is depicted in Figure 2.9. A neural network comprises of neurons with parameters $\Theta = \{\mathbf{W}, \mathbf{B}\}$, where $\mathbf{W}$ is a set of weights and $\mathbf{B}$ a set of biases. At each layer, the input image $\mathbf{X}$ is convolved with a set of K kernels $\mathbf{W} = \{\mathbf{W_1}, \mathbf{W_2}, ..., \mathbf{W_K}\}$ with biases $\mathbf{B} = \{\mathbf{b_1}, \mathbf{b_2}, ..., \mathbf{b_K}\}$ added to generate new feature X_k. These features then are subjected to an element-wise non-linear transform by an activation function $\sigma(.)$, which introduces a non-linearity. ReLU are widely used for this purpose (for faster training), while a softmax is used at the end of the fully connected layer [90]. This process then is repeated for every convolutional layer l:

$$\mathbf{X_k^l} = \sigma(\mathbf{W_k^{l-1}} * \mathbf{X^{l-1}} + b_k^{l-1})$$

Convolutional layers provide activation maps, which then are downsampled by pooling layers. In pooling layers, the pixel values of neighborhoods are aggregated, typically by the max or mean operation. Hence, the amount of parameters in the network is reduced. At the end of the convolutional stream of the network, fully-connected layers are added. Here, weights are

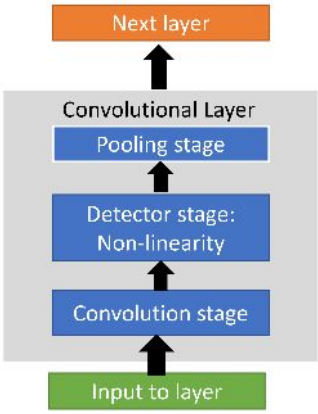


Figure 2.8: The convolutional network is viewed as a small number of relatively complex layers, with each layer having many stages. (The figure is adapted from [57])

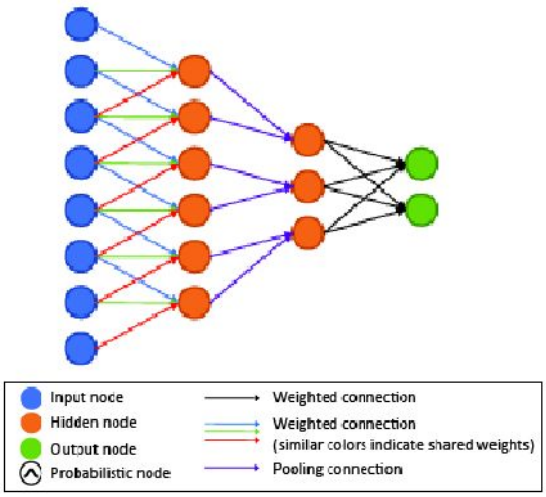


Figure 2.9: Node graphs of 1D representations of CNN architectures (the figure is adapted from [90])

no longer shared. At the final layer of the network the activations are mapped to a distribution over classes through the softmax function:

$$P(y|\mathbf{x};\Theta) = softmax(\mathbf{x};\Theta) = \frac{e^{\mathbf{w}_i^T\mathbf{x}+b_i}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T\mathbf{x}+b_k}}$$

Here w_i indicates the weight vector leading to the output node associated with class i. Due to non-convex nature of commonly used objective functions, CNNs are trained iteratively to fit the parameter Θ to a dataset D.

The most common way to train a neural network is by back-propagation algorithm. The back-propagation algorithm is used to learn the optimal set of model parameters (weights or coefficients) for the convolutional filters. The updates on each layer parameters are computed by back-propagating the error from the loss function. End-to-end learning and back-propagation also reduce the dependency of on hand-crafted image features extraction and pre-processing techniques.

Mini-batch training is a recommended variant of gradient descent for deep learning (instead of full dataset). Mini-batches provide tolerance to noisy data and reduces the chance of the optimization getting stuck in local minima. Additionally, iterative mini-batch update scheme enables the use of large scale datasets for training. Batch size is one of the parameters to be tuned in the learning process. Small values give a learning process that converges quickly at the cost of noise in the training process. Large values give a learning process that converges slowly with accurate estimates of the error gradient. In practice, the update scheme can be seen as to minimizing the negative log-likelihood:

$$\mathbf{arg}\min_{\Theta} -\sum_{n=1}^{N} \mathbf{log}[P(y_n|\mathbf{x}_n;\Theta)]$$

As negative log-likelihood is equivalent to minimizing the cross entropy, this results in the categorical cross-entropy for multi-class tasks.

In CNN training, tuning hyperparameters can help squeeze out extra performance from a

network. Some of the hyperparameters are:

- Learning rate: Controls how much to update the weight in the optimization algorithm.
- Number of epoch: Number of times the entire training set pass through the neural network
- Batch size: Mini-batch is usually preferable in the CNN learning process to achieve a relatively quicker and more robust convergence.
- Activation layer: It introduces non-linearity to the model, usually rectified linear units (ReLU) works well with CNN.
- Number of hidden layer: Adding more layers is better to reduce test error, but increases computational cost to train the network.
- Weight of initialization: Initialize the weights with small random numbers to prevent dead neurons, but not too small to avoid zero gradient.
- Dropout rate: Preferable regularization technique to avoid overfitting in deep neural networks.

CNNs have dominated the medical image analysis in the last years since they can learn complex and task-specific abstract representations [90]. Moreover, they can scale to large training images well with strong representation power. Several CNN-based methods have been served as the current state-of-the-art for medical image segmentation, such as U-Net [115], FCN[92] and DeepMedic [74].

2.2 Transfer Learning

Although machine learning has reached significant success, most of the techniques only work well when the training and testing data are drawn from the same distribution. When the distributions are different, the models need to be retrained using newly collected training data, which is expensive and time consuming. Hence, *transfer learning* between domains can reduce the

effort of retraining or recollect the training data. Transfer learning aims to transfer knowledge between related source and target domains.

In medical images, the mismatch of training and test datasets can include factors such as different scanners, type of sequences, or modality. This often causes algorithms to perform poorly across different domains. Generally, transfer learning is a framework to address this domain disparity problem. In a scenario where the source and target tasks are the same but the source and target domains are different, domain adaptation is a specific area of transfer learning to solves this problem [108].

Domain adaptation methods have been studied to generalize learned models [109]. The domain of labeled training data is known as the source domain, with the test dataset as the target domain. In general, domain adaptation can be categorized into three settings, supervised, semi-supervised, and unsupervised. We will explain the differences between the methods below.

2.2.1 Supervised Domain Adaptation

We define supervised domain adaptation as having a large amount of *labeled* data in the source domain, a large amount of *unlabeled* in the target domain, and additionally a small budget for acquiring labels in the target domain [120]. Here, labels from the target domain are used in the process.

An approached by [55] explored a supervised domain adaptation method by introducing a weighting scheme in Random Forests and Support Vector Machine (SVM) for segmentation. This approach is found to preserve the segmentation quality even though only sparse annotated training data used. Since the labelling time of novel training data can be reduced, it allows for a more efficient adaptation to variations in scanners, modalities or protocols.

Another approach proposed in [143] tackles the difficulty of segmentation of images from different scanners and imaging protocols. They investigated the use of SVM-based transfer classifiers in transfer learning. Their experiments show that the use of transfer classifiers reduce

the required number of labeled samples from the target domain, and outperform common supervised-learning approaches.

In the context of CNNs, image representations learned with CNNs on large-scale annotated datasets can be efficiently transferred to other visual recognition tasks with limited amount of training data [107]. The key idea of this work is that features extracted from the CNN can act as a generic extractor of image representation [124], which can be pre-trained on the source task and then re-used on other target tasks. Training a CNN from a set of pre-trained weights is commonly called *fine-tuning* and has been used successfully in several applications [124, 133]

CNNs can learn transferable features that generalise to novel tasks for supervised domain adaptation. The early layers of a CNN learn low level image features but the late layers learn high-level features, which are specific to a given task [163]. Therefore, fine-tuning the last few layers is usually sufficient for transfer learning. However, fine-tuning a deep network may not be efficient when the distance between the source and target applications is significantly different. One may need to fine-tune the early layers as well. We refer to tuning the last few convolutional layers as 'shallow tuning' and tuning all the convolutional layers as 'deep tuning' [133].

Several strategies have been proposed to make fine-tuning between different domains be more efficient. The work by [137] proposed Deep Domain Confusion (DDC), which utilizes the Maximum Mean Discrepancy (MMD) measure as an additional loss function for the fine-tuning step to adapt the last fully connected layer. Meanwhile, Deep Adaptation Network (DAN) [93] fine-tunes, not only the last fully connected layer, but also some convolutional and fully connected layers underneath. This has been shown to outperform DDC.

2.2.2 Semi-supervised Domain Adaptation

We refer to semi-supervised domain adaptation as having labeled data in source, and both labeled and unlabeled data in target. The difference is in semi-supervised learning, the goal is to use the unlabeled samples to improve the performance of a classifier [26]. Commonly, semi-

supervised approaches work by exploiting an additional assumption from the labeled data, for example by iterating a classifier to find best possible labels for unlabeled samples and retraining itself.

There have been several works in computer vision that explore semi-supervised domain adaptation, ranging from the utilization of a classifier to dictionary learning [109]. In the literature, semi-supervised techniques based on SVMs been proposed, such as DASVM (Domain Adaptation Support Vector Machine), uses an iterative method to gradually adapt the discriminant function by adding the unlabeled samples to the training set [18]. Another existing approach, based on SVM, include the use of adding unlabeled instance similarity constraints to labeled examples [38] and projections of new representation from source and target domains into a subspace to reduce the distribution mismatch [152].

2.2.3 Unsupervised Domain Adaptation

We define unsupervised domain adaptation as having labeled data in only source domain, but labeled training set from target domain is not available, which is often the case in clinical practice and in multi-center studies. Unsupervised domain adaptation has gained much attention recently with the advancement of deep learning, which can discover good representations structure in the input distribution [11]. We consider different techniques of unsupervised domain adaption in the literature, ranging from classifier optimization [46, 73] to invariant feature representation [137, 46].

Previous studies on unsupervised domain adaptation focuses on minimizing the distance between domains in feature space. For example, the MMD was minimized together with a task-specific loss to learn domain invariant and semantic-meaningful features with single [137] or multiple layers [93]. Lempitsky et al. [46] proposed to extract domain-invariant features by sharing weights between two classifiers. The method measures and minimizes the discrepancy between feature distributions for the two domains, by augmenting gradient reversal layer and training using standard back-propagation. This gradient reversal strategy can minimize the

loss of the label classifier and maximize the loss of the domain classifier. In the spirit of multi-task learning, DRCN (Deep Reconstruction-Classification Network) [52] simultaneously finds a good label while learning the structure of the target images. They show that auxiliary learning tasks can help the main task learn better by encoding the useful information from the target domain.

Auto-encoders are commonly used as a pre-training model for deep learning, since it is very good in domain generalization task with its ability to reconstruct input. Several models exploit auto-encoder as a way of unsupervised domain adaptation. These models include MTAE (Multi-task Autoencoder) [51], which involves multi-task feature learning shared across domains and bi-shifting auto-encoder [75] to shift the samples between source and target domains so that their distributions can be matched. However, such efforts in learning domain-invariant features in the context of shallow neural networks have limited capacity to surpass deep CNN.

Alternatively, with the emergence of Generative Adversarial Network (GAN) [58], the mapping between domains can be implicitly learned via the adversarial loss. Later, [14] adapts the source-domain images to appear as if they are drawn from target domain by learning a transformation in the pixel space via GAN based approach. Adversarial neural network, applied in medical imaging [73], can generalize well across differernt modalities in brain MRI. However, optimizing classifiers and changing architectures are found to be a bit tricky, and thus require more work.

Prior works have tried classifier adaption to optimise learning between domains. Residual transfer networks (RTN) [94] have been proposed to simultaneously learn transferable features and adaptive classifiers. In feature adaption, they add bottleneck on deep feature layers so that the feature distribution between source and target made similar. In classifier adaptation, RTN bridges the supervised classifier by assuming that the source and target classifiers differ by a small residual function. By combining both (feature and classifier adaptation), they achieve an effective domain adaptation.

Later, the work by [39] uses the domain adaptation module (DAM) to map the target input to features aligned with source feature space. To discriminate the feature space of both domain, a domain critic module (DCM) is set up. Moreover, they optimise DAM and DCM with

an adversarial loss. This combination of elaborated techniques has shown to work on cross-modality unsupervised domain adaptation.

2.3 Multi-modal Learning

Humans learn the world through various modalities: we see things, hear sounds, touch, smell odors, and taste flavors. In this context, multi-modal refers to the involvement of multiple modalities [129]. In order for machine learning to understand the world, the machine needs to be able to interpret multi-modal signals together. Multi-modal machine learning aims to build a model that can relate the correspondence between the heterogeneity of data and inferring meaningful information from multiple modalities. However, learning between modalities faces several challenges [9], such as the representation of multi-modal data, translation among modalities, alignment from multiple modalities, fusion of multi-modal information, and transfer knowledge between modalities [9].

Recently, deep learning has become a very popular method to explore multi-modal learning. Deep learning based multi-modal learning has shown to produce useful joint representation between modalities [151]. According to [151], there are two main approaches that have been used for deep-learning based multi-modal learning: representation learning based and Canonical Correlation Analysis (CCA). The first approach learns a joint representation from input modalities, which includes Deep Autoencoder (DAE) models. The second approach learns separate representations from modalities while maximizing their correlation.

A work in multi-modal in deep learning has been addressed by [101] to learn features over multiple modalities. Their bimodal auto-encoder structure relates the 'mid-level' correlation between pixels from lips video and audio waveform to classify speech. In computer vision tasks, multi-modal learning had gained an interest as prior works have shown a better performance [151, 146, 131].

2.3.1 Learning on Paired Images

In computer vision, prior works have explored learning between images and texts [131], and video and audio [101]. Although each modality has a different dimension and structure, the data are paired. Paired data means each modality has a one-to-one relationship between values in the two data sets from different modalities. To uncover the high nonlinear relation between each modality and the joint representation, deep learning based multi-modal architectures are applied. To learn joint representations, the common approach is to share the last layers as modality-specific networks [131, 101].

In medical imaging domain, most of the approaches in literature explore multi-modal segmentation on paired images obtained from a single imaging-modality, which is MRI. A patient may have corresponding T1, T1w, T2w, and FLAIR images from an MRI scan [74, 62, 42, 61]. All of the aforementioned previous works on multi-modal segmentation uses paired data, meaning that the images are acquired from the same subject and registered between different modalities. The simplest, yet the most common approach to those multi-modal segmentation problems, is by implicitly fusing multi-modal information in the network during training without caring about which modalities. Each modality is treated as one input channel, which is then processed in the corresponding architecture [74, 61].

2.3.2 Learning on Unpaired Images

There are few works in learning on unpaired images, as most existing multi-modal learning algorithm assume access to paired data for training. However, in the real practice, paired set of data is rarely available. In computer vision domain, learning from unpaired images is generally seen on image translation problem [168], image captioning[99], and image recognition [80].

One popular learning approach, of using unpaired samples, is image translation using adversarial networks. Cycle-GANs [168] are the most common approach to tackle unpaired image-to-image translation, in cases where paired training data is not available. With Cycle-GAN, the mapping of an image from one domain to another domain could be estimated by satisfying

marginal matching (using GAN) and cyclic-consistency. Cycle-GAN has been used for many creative applications in the computer vision domain.

Meanwhile, image captioning for unpaired multi-modal data has been studied by [99]. Matching natural language descriptions for images/videos has recently become a popular topic in cross-modal learning in the last few years. Using Deep Matching Autoencoders (DMAE), it simultaneously optimises parameters of representation learning in auto-encoders with the pairing of unpaired multi-modal data.

The application of scene recognition using unpaired multi-modal datasets has been explored in [80]. Using multi-modal encoder-decoder networks, the method utilizes auxiliary unpaired data to train self-encoding paths and to consequently improve the cross-modal performance. Their multi-modal input consists of paired and unpaired RGB images, depth, and semantic labels, for multi-task scene understanding.

In medical imaging, learning unpaired images between modalities have shown to be useful in cross-modality synthesis tasks. A work by [67] exploits unpaired images from different domains in MRI scans by joint convolutional sparse coding with hetero-domain image alignment to minimize the dissimilarity between the feature distributions. A more challenging image synthesis between CT and MR has been explored by [70] using both paired and unpaired images. They proposed a CycleGAN-like approach but with dual cycle-consistent terms (forward and backward) for paired and unpaired data. The approach is shown to be superior for both paired and unpaired training data.

2.4 Population Image Analysis

In this era of big data, population studies with large-scale image data pose a new challenge. UK Biobank [130] is one of the largest ongoing population studies that collects several types of imaging data, such as; 3T MR imaging of the brain, 1.5T MR imaging of the heart and body, DXA images of bone and joint, and carotid ultrasound images. With the availability of such vast datasets, large-scale population image analysis offers a great opportunity for a

better understanding of disease pathology, data visualization, and data interpretation [111]. However, numerous challenges exist, such as image quality assurance, storing large scale of data, managing computing resources for analysis, and adapting existing algorithms for robust large-scale analysis.

Biomarker identification for a disease, derived from large population studies, may reveal important knowledge for disease diagnosis. Imaging biomarkers are useful for clinical assessment and drug delivery development. Machine learning techniques, such as manifold learning, is one of the techniques that can be used to extract useful biomarkers [60]. Once the biomarkers are identified, the population can be clustered and grouped according to their health-related outcomes. Neighbourhood Approximation Forest (NAF) is one of the neighbourhood-based methods that can be applied to infer biomarkers from imaging database, such as the ages in the population from brain MRI dataset [79]. NAF is a variant of random-forest that allows automatic feature selection for approximating neighbourhood structures.

The shape of an organ can also be a useful biomarker in population study. Anatomical and pathological organ shape variations are important for disease diagnosis. Dissimilarity between shapes can be computed once the shape descriptors are defined. A common shape descriptor is based on the eigen-system of Laplace operators called spectral signatures. A new shape distance called Weighted Spectral Distance for Measuring Shape Dissimilarity (WESD) provides a consistent and scale-invariant measurement based on eigenvalues [77]. Figure 2.10 shows our 3D visualisation of an organ (liver) in varying shapes from MALIBO dataset using MayaVi [113] in Python.

2.4.1 Quality Control

Quality control is an initial stage before medical images are analysed for high-level interpretation. This task can be time-consuming when dealing with large-scale imaging studies where hundreds or thousands of subjects are scanned. Experts must visually inspect them to ensure that the images are complete, free of artifacts and are clinically usable. However, this process is costly and prone to error since the quality assessment criteria is commonly subjective.

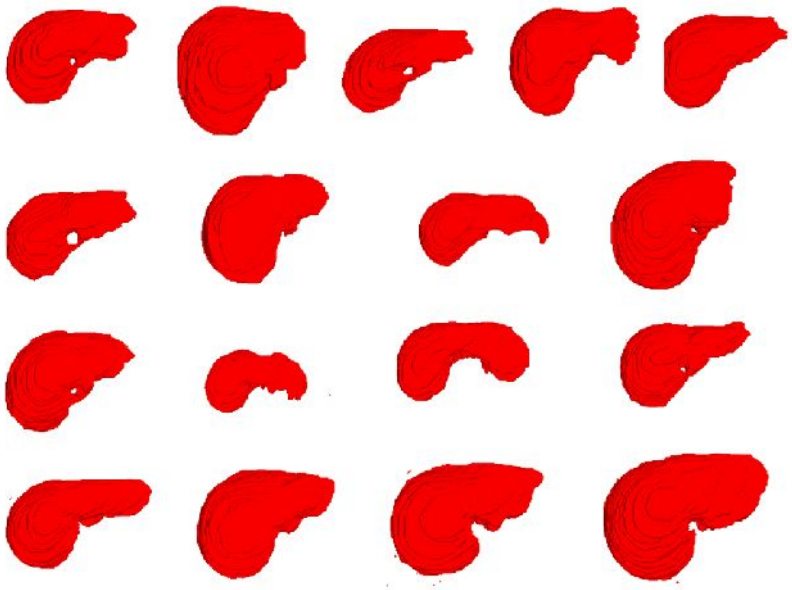


Figure 2.10: Varying organ shapes in the population is shows in rendered 3D livers from MALIBO dataset

Large scale imaging studies, such as UK Biobank dataset face particular quality control challenges [130]. UK Biobank whole-body MR images are acquired using the Dixon method that produces scans with four types of images: in-phase, out-of-phase, fat-only, and water-only. The imaging data is acquired using six scanning stations for each subject. Then, the images are stitched together to produce a volume covering from shoulders to knees. An automated quality control framework for whole-body MRI is shown in Figure 2.11.

For whole-body MRI scans, common quality control issues are listed below:

Missing stations

If a subject does not have the complete six stations of Dixon scans, then it should be rejected by the automated quality control system since it cannot form a complete stitched whole-body volume.

Repeated stations

In whole-body MRI scans, each station represents a unique body region. Station 1 gives shoulder and part of the heart. Station 2, in majority, shows the liver. Station 3 contains bowels and

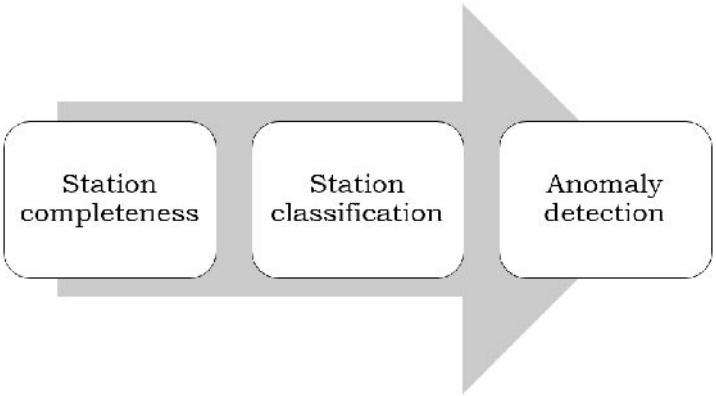


Figure 2.11: Automated quality control stages

some abdominal organs (kidneys, spleen). Station 4 points the bladder and pelvis. Station 5 locates both thighs. Station 6 gives the appearance of legs until knees. Correct scanned images should cover a continuous scan from shoulders to knees. An example case of repeated station is given in Figure 2.12. The content of supposed-to-be-station 2 is repeated in Station 1, so that the stitched volume has seven stations instead of six. Repeated station leads to wrong content in all stations as the content of subsequent stations is shifted.

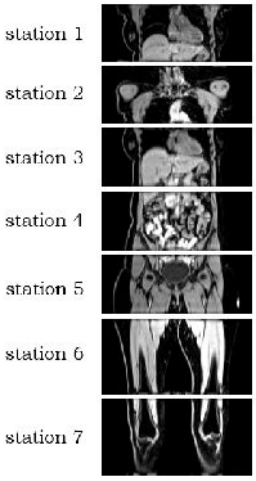


Figure 2.12: Repeated station leads to wrong station content

A machine learning implementation to solve this problem is station classification, which aims recognise the captured body regions. Content-Based Image Retrieval (CBIR) is known to tackle this problem, by retrieving the most visually similar images to a given query image from an

image database. However, the selection of similarity metric affects the performance of CBIR. The work of [158] uses Mutual Information (MI) as the similarity measure for CBIR.

MI between two images X and Y can be defined as:

$$MI(X,Y) = \sum_{x \in U} \sum_{y \in V} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{2.5}$$

Here X is a random variable describing the imaging intensities with a statistical distribution $p(x)$, similarly for Y [158]. Mutual information can also be written as:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.6}$$

with

$$H(X) = -\sum_{x \in U} p(x) log(p(x)) \tag{2.7}$$

Here $H(X)$ and $H(Y)$ are the entropy of images X and Y, and $H(X,Y)$ is the joint entropy of X and Y. In information theory, entropy is a measure of unpredictability of information content. Equation 2.6 shows that mutual information $MI(X,Y)$ is defined as the relative entropy between the joint distribution $p(x,y)$ and the product $p(x)p(y)$ of the marginal distributions. From the relationship between MI and entropy, we can infer that mutual information $MI(X,Y)$ can be interpreted as the reduction of the uncertainty of X due to knowledge of Y.

In the station classification problem, X is the query station while Y is the reference station. As illustrated in Figure 2.13, MI is computed between the query station and the six-stations in the reference, where the highest MI score gives the matching station with the query content.

Fat-water Swaps

The Dixon method often suffers from fat-water swaps due to phase wrapping and local minima in the optimization [37]. The artifact of fat-water inversion appears quite often and affects up to 10% of routinely acquired Dixon images. This may lead to severe problems for any subsequent analysis. In whole-body scans, there are two types of fat-water swaps:

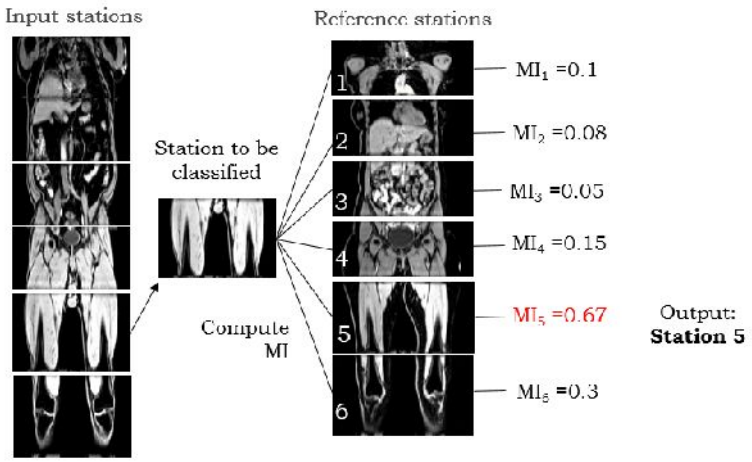


Figure 2.13: Station Classification: Calculate MI between a query station and each of six-stations in reference. The highest MI score gives the right station name.

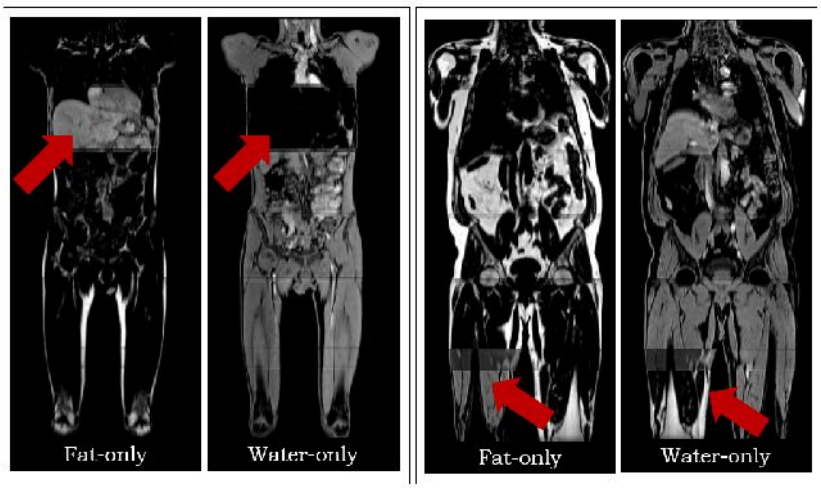


Figure 2.14: Inter-station (Left) and intra-station (Right) fat-water swap cases

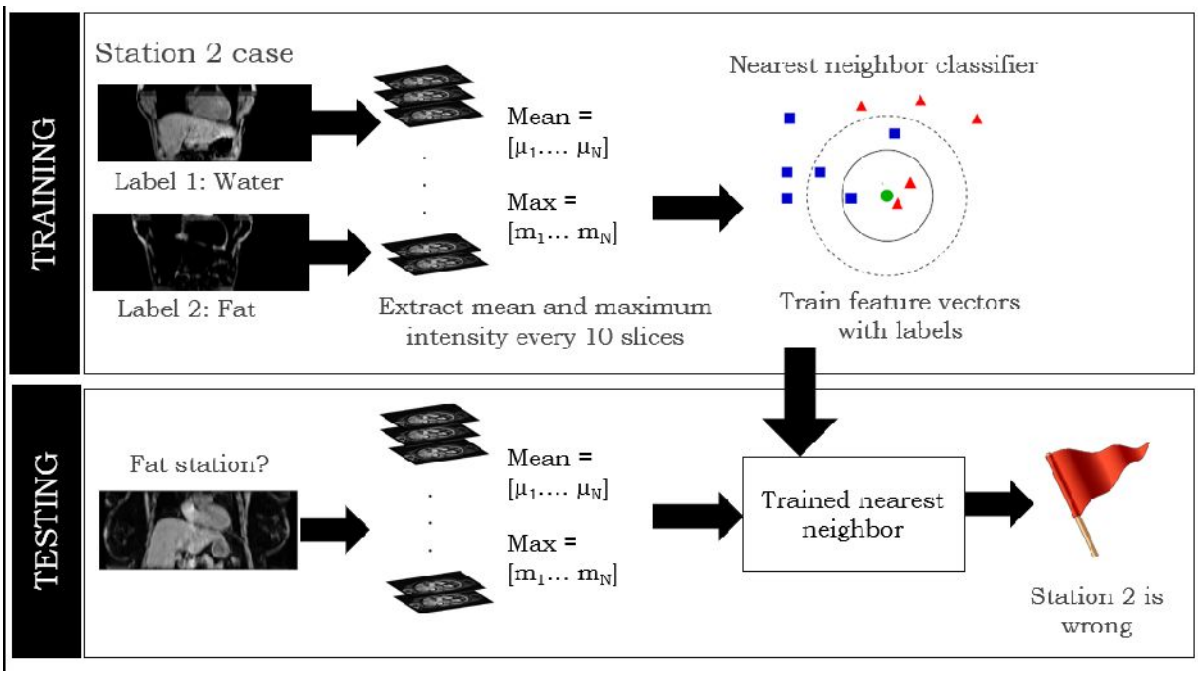


Figure 2.15: Anomaly detection for each station

Inter-station fat-water swap : A case where the whole content in a station is swapped, as shown in Figure 2.14. This kind of fat-water swap can be easily fixed manually by physicians; by swapping and renaming the stations.

Intra-station fat-water swap: A case where only part of a station is swapped, as we can see in Figure 2.14 left; one leg in station 6 is swapped. In other cases of intra-station fat-water swaps, we also found cases that only swap an organ, for example liver in fat-image is swapped with liver from water-image. This type of fat-water swap is hard to be fixed and may require the patients to be rescanned.

Fat-water swap causes the stations to behave and appear differently from the 'normal' stations. An anomaly detection scheme [23] based on semi-supervised method is one approach to tackle this problem. Features are extracted for a k-Nearest Neighbour classifier. The nearest neighbours stores instances of training data and learns them based on the number of neighbours specified. The classifier learns the normal distribution of water and fat images for every station. Then, during testing, it assigns the data to the class that is the most representing within the nearest neighbours (Figure 2.15). Normal stations should be close neighbors, while an abnormal station is classified as another label (outlier) since it is further away from its neighbours.

A more sophisticated approach has been proposed to detect and to correct the fat-water swaps using Dixon-Fix [54]. Given the in-and out-of-phase images, the conditional distribution of fat-and-water only images appearance could be learned via Random Forest. Graph cuts are then used to estimate the swap-labeling. The fat-water swaps in all cases can be successfully corrected, while still leaving a challenge in swap-free images (false negative).

2.4.2 Manifold Learning for Population Study

Manifold learning is one of the unsupervised machine learning techniques for dimensionality reduction. Unlike Principle Component Analysis (PCA) [96] and Multi Dimensional Scaling (MDS) which are linear dimensionality reduction techniques, manifold learning consists of nonlinear dimensionality reduction techniques.

Manifold learning has been applied in various problems in medical imaging, including multi-atlas segmentation, registration, motion modelling, population study, biomarker discovery, and classification [3]. Some of the commonly used non-linear dimensionality reduction techniques include ISOMAP (Isometric mapping), Locally Linear Embedding (LLE), Laplacian Eigenmaps, and t-Distributed Stochastic Neighbor Embedding (t-SNE) [95].

The classic ISOMAP approach seeks to learn a lower-dimensional embedding that maintains geodesic distances between all points. Geodesic distance is the distance between two points measured over the manifold. LLE and Laplacian Eigenmaps are sparse spectral techniques. Both construct a low-dimensional representation by solving an eigen decomposition of the affinity matrix. The advantages of using the sparse spectral techniques is the local structure of the data is retained. The most recently proposed technique, t-SNE, converts data affinities to probabilities. While ISOMAP, LLE and Laplacian Eigenmaps are best suited to unfold a single continuous low dimensional manifold, t-SNE is particularly well suited for the visualization of high-dimensional datasets. However, since t-SNE uses Kullback-Leibler (KL) divergence for solving the embedding, the result is not convex and will produce different local minima for different initializations [141]..

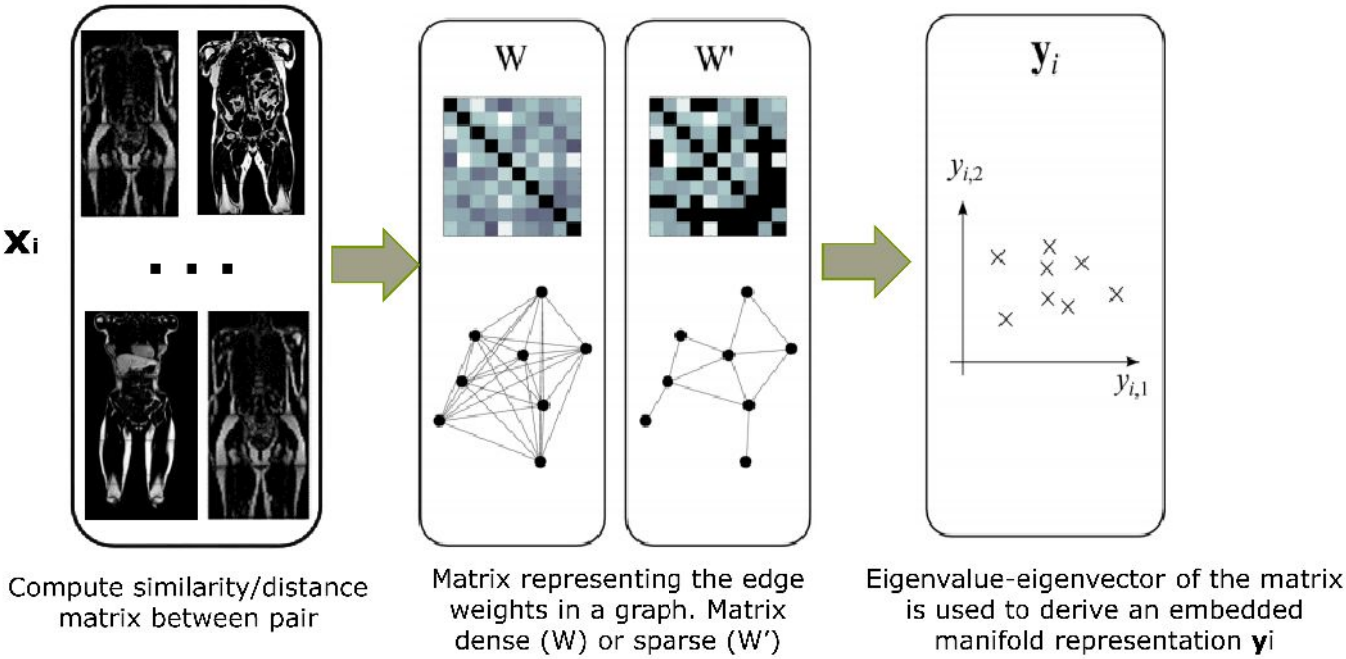


Figure 2.16: Overview of manifold learning technique: representing data in lower dimension by embedded manifold representation

Figure 2.16 shows an overview of manifold learning techniques. A full data $\mathbf{X}$ consists of a set of N images $x_1, ..., x_N$ where D is the number of voxels per image x. It is possible to learn from $\mathbf{X}$, a new low dimensional ($d << D$) representation of the images $\mathbf{Y} = y_1, ..., y_N$. Using manifold learning, the similarity/distance matrix is used to represent the relation between pair of images. This matrix can be viewed as a graph where each node is an image and the weight of each edge is the similarity/distance between image pair. Some manifold learning methods use the complete graph (dense) or sparser representation (smaller number of edges, restricted to local neighborhood) to model relation among the data.

In LLE, local properties are constructed by writing the data point as a linear combination of weights of its k nearest neighbors. In Laplacian Eigenmaps, local properties are based on the pairwise distance between nearest neighbours [141]. Any practical decisions to chose between the many techniques of manifold learning should be based on their objective functions.

The application of manifold learning on a population study of neonatal brain development shows that the same patterns of morphology are clustered together [3]. Image and non-image (metadata) information, such as measurement of gender, age, genotype, can also be combined

in manifold learning to extract more accurate biomarkers, and produce classification from low-dimensional space [148]. Bengio et al. [12] presented an algorithm to obtain an embedding for a new data (out-of-sample) point on spectral-based manifold learning techniques, such as MDS, ISOMAP, LLE, and Laplacian Eigenmaps.

Chapter 3

Learning to Predict the Segmentation Accuracy

This chapter is based on:

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse classiffication accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*, 36(8):1597-1606, 2017.

From the previous chapter, we learn that quality control on automatic segmentation results is important, especially in a large imaging database. The ability to assess the level of accuracy on new data, and in particular, to detect when an automatic segmentation method fails is crucial in clinical practice. However, this is difficult to achieve due to absence of ground truth. Segmentation accuracy on clinical data might be different from what is found through cross-validation because validation data is often used during incremental method development, which can lead to overfitting and unrealistic performance expectations. Before deployment, the performance is quantified using different metrics, for which the predicted segmentation is compared to a reference segmentation, often obtained manually by an expert. But little is known about the real performance after deployment when a reference is unavailable.

In this chapter, we introduce the concept of *reverse classification accuracy* (RCA) as a frame-

work for predicting the performance of a segmentation method on new data. In RCA we take the predicted segmentation from a new image to train a reverse classifier which is evaluated on a set of reference images with available ground truth. The hypothesis is that if the predicted segmentation is of good quality, then the reverse classifier will perform well on at least some of the reference images. We validate our approach on multi-organ segmentation with different classifiers and segmentation methods. Our results indicate that it is indeed possible to predict the quality of individual segmentations, in the absence of ground truth. Thus, RCA is ideal for integration into automatic processing pipelines in clinical routine and as part of large-scale image analysis studies.

3.1 Introduction

Segmentation is an essential component in many image analysis pipelines that aim to extract clinically useful information from medical images to inform clinical decisions in diagnosis, treatment planning, or monitoring of disease progression. A multitude of approaches have been proposed for solving segmentation problems, with popular techniques based on graph cuts [15], multi-atlas label propagation [68], statistical models [65], and supervised classification [49]. Traditionally, performance of a segmentation method is evaluated on an annotated database using various evaluation metrics in a cross-validation setting. These metrics reflect the performance in terms of agreement [173] of a predicted segmentation compared to a reference or ground truth (GT). For simplicity, we use the term ground truth to refer to the best known reference, which is typically a manual expert segmentation. Commonly used metrics include Dice's similarity coefficient (DSC) [36] and other overlap based measures [34], but also metrics based on volume differences, surface distances, and others [132, 35, 86]. A detailed analysis of common metrics and their suitability for segmentation evaluation can be found in [78].

Once a segmentation method is deployed in routine, we may still unsure about its real performance on new data. Due the absence of GT, it is not possible to assess performance using traditional evaluation measures. However, it is critical to be able to assess the level of accuracy

on clinical data [142], and in particular, it is important to detect when an automatic segmentation method fails. Especially when the segmentation is an intermediate step within a larger automated processing pipeline where no visual quality control of the segmentation results is feasible. This is of high importance in large-scale studies such as the UK Biobank Imaging Study [130] where automated methods are applied to large cohorts of several thousand images, and the segmentation is to be used for further statistical population analysis.

In this chapter, we are asking the question whether it is possible to assess segmentation performance and detect failure cases when there is no GT available to compare with. One possible approach to monitor the segmentation performance is to occasionally select a random dataset, obtain a manual expert segmentation and compare it to the automatic one. While this can merely provide a rough estimate about the average performance of the employed segmentation method, in clinical routine we are interested in the per case performance and want to detect when the automated method fails. The problem is that the performance of a method might be substantially different on clinical data, and is usually lower than what is found through cross-validation on annotated data carried out beforehand due to several reasons. First, the annotated database is normally used during incremental method development for training, model selection and fine tuning of hyper-parameters. This can lead to overfitting [40] which is a potential cause for lower performance on new data. Second, the clinical data might be different due to varying imaging protocols or artefacts caused by pathology. To this end, we propose a general framework for predicting the real performance of deployed segmentation methods on a per case basis in the absence of GT.

3.1.1 Related Work

Retrieving an objective performance evaluation without GT has been an issue in many domains, from remote sensing [10], graphics [91], to marketing strategies [20]. In computer vision, several works evaluate the segmentation performance by looking at contextual properties [29], by separating the perceptual salient structures [48], or by automatically generating semantic GT [56, 88]. However, these methods cannot be applied to a more general task, such as an

image with many different class labels to be segmented. An attempt to compute objective metrics, such as precision and recall with missing GT is proposed by [81] but it cannot be used for data sets with partial GT since it applies a probabilistic model under the same assumptions. Another stand-alone method is to consider a meta-evaluation framework, where image features are used in a machine learning setting to provide a ranking of different methods [156]. But this does not allow the estimation of segmentation performance on an individual image level.

Meanwhile, unsupervised methods [22, 157] aim to estimate the segmentation accuracy directly from the images and labelmaps using, for example, information-theoretic and geometrical features. While unsupervised methods can be applied to scenarios where the main purpose of segmentation is to yield visually consistent results that are meaningful to a human observer, the application in medical settings is unclear.

When there are multiple reference segmentations available, a similarity measure index can be obtained by comparing an automatic segmentation with the set of references [139]. In medical imaging, the problem of performance analysis with multiple references which may suffer from intra-rater and inter-rater variability, has been addressed [147, 89]. The STAPLE approach [147] has lead to the work of Bouix et al. [13] that proposed techniques for comparing the relative performance of different methods without the need of GT. Here, the different segmentation results are treated as plausible references, thus can be evaluated through STAPLE and the concept of common agreement. Another work by [127] has quantitatively evaluated the performance of several segmentation algorithms by region-correlation matrix. The limitation of this work is that it cannot evaluate the segmentation performance of a particular method on a particular image independently.

Recent work has explored the idea of learning a regressor to directly predict segmentation accuracy from a set of features that are related to various segmentation energy terms [76]. Here, the assumption is that those features are well suited to characterise segmentation quality. In an extension for a security application, the same features as in [76] are extracted and used to learn a generative model of good segmentations that can be used to detect outliers [59]. Similarly, the work of [45] considers training of a classifier that is able to discriminate between

consistent and inconsistent segmentations. However, the approaches [76, 45] can only be applied when a training database with good and bad segmentations is available from which a mapping from features to segmentation accuracy can be learned. Examples of bad segmentations can be generated by altering parameters of automatic methods, but it is unclear whether those examples resemble realistic cases of segmentation failure. The generative model approach in [59] is appealing as it only requires a database of good segmentations. However, there is still the difficulty of choosing appropriate thresholds on the probabilities that indicate bad or failed segmentations. Such an approach cannot be used to directly predict segmentation scores such as DSC, but can be useful to inform automatic quality control or to automatically select the best segmentation from a set of candidates.

In the general machine learning domain, the lack-of-label problem has been tackled by exploiting transfer learning [162] using a reverse validation to perform cross-validation when the number of labeled data is limited. The basic idea of reverse validation [162] is based on reverse testing [41], where a new classifier is trained on predictions on the test data and evaluated again on the training data. This idea of reverse testing is closely related to our approach (but ours can predict the performance of a classifier in a per case basis) as we will discuss in the following.

3.1.2 Contribution

The main contribution of this chapter is the introduction of the concept of *reverse classification accuracy* (RCA) to assess the segmentation quality of an individual image in the absence of GT. RCA can be applied to evaluate the performance of any segmentation method on a per case basis. To this end, a classifier is trained using a single image with its predicted segmentation acting as *pseudo* GT. The resulting *reverse classifier* (or RCA classifier) is then evaluated on images from a reference database for which GT is available. It should be noted that the reference database can be the training database that has been used to train, cross-validate and fine-tune the original segmentation method. The assumption is that in machine learning approaches, such a database is usually already available, but it could also be specifically constructed for the purpose of RCA.

Our hypothesis is that if the segmentation quality for a new image is high, then the RCA classifier trained on the predicted segmentation used as pseudo GT will perform well at least on some of the images in the reference database, and similarly, if the segmentation quality is poor, the classifier is likely to perform poorly on the reference images. For the segmentations obtained on the reference images through the RCA classifier, we can quantify the accuracy, e.g., using DSC, since reference GT is available. It is expected that the maximum DSC score over all reference images correlates well with the real DSC that one would get on the new image if GT were available. While the idea of RCA is similar to reverse validation [162] and reverse testing [41], the important difference is that in our approach we train a reverse classifier on every single instance while the approaches in [162, 41] train single classifiers over the whole test set and its predictions jointly to find out what the best original predictor is. RCA has the advantage of allowing to predict the accuracy for each individual case, while at the same time aggregating over such accuracy predictions allows drawing conclusions for the overall performance of a particular segmentation method.

In the following, we will first present the details of RCA and then evaluate its applicability to a multi-organ segmentation task by exploring the prediction quality of different segmentation metrics for different combinations of segmentation methods and RCA classifiers. Our results indicate that, at least to some extent, it is indeed possible to predict the performance level of a segmentation method on each individual case, in the absence of ground truth. Thus, RCA is ideal for integration into automatic processing pipelines in clinical routine and as part of large-scale image analysis studies.

3.2 Reverse Classification Accuracy (RCA)

The RCA framework [1] is based on the idea of training reverse classifiers on individual images utilizing their predicted segmentation as pseudo GT. An overview of the RCA framework is shown in Fig. 6.1. In this work, we employ three different methods for realizing the RCA classi-

[1] Source code available at `https://github.com/vanya2v/RCA`

fier and evaluate each in different combinations with three state-of-the-art image segmentation methods. Details about the different RCA classifiers are provided in the following.

3.2.1 Learning Reverse Classifiers

Given an image I and its predicted segmentation S_I, we aim to learn a function $f_{I,S_I}(\vec{x}) : \mathbb{R}^n \mapsto C$ that acts as a classifier by mapping feature vectors $\vec{x} \in \mathbb{R}^n$ extracted for individual image points to class labels $c \in C$. In theory, any classification approach could be utilized within the RCA framework for learning the function f_{I,S_I}. We experiment with three different methods reflecting state-of-the-art machine learning approaches for voxel-wise classification and atlas-based label propagation.

Atlas Forests The first approach we consider for learning a RCA classifier is based on the recently introduced concept of Atlas Forests (AFs) [170] which demonstrates the feasibility of using Random Forests (RFs) [16] to encode individual atlases, i.e., images with corresponding segmentations. Random Forests have become popular for general image segmentation tasks as they naturally handle multi-class problems and are computationally efficient. Since they operate as voxel-wise classifiers, they do not (necessarily) require pre-registration of the images neither at training nor testing time. Although in [170] spatial priors have been incorporated by means of registering location probability maps to each atlas and new image, this is not a general requirement for using AFs to encode atlases. In fact, the way we employ AFs within our RCA framework does not require any image registration.

The forest-based RCA classifiers in this work are trained all with the same set of parameters of maximum depth 30 and 50 trees. As we follow a very standard approach for RFs, we refer to [33, 170] for more details. Worth to note that, similar to previous work, we employ simple box features which can be efficiently evaluated using integral images. This has the advantage that feature responses do not need to be precomputed. Instead, we randomly generate a large pool of potential features (typically around 10,000) by drawing randomly values for the feature parameters such as box sizes and offsets from predefined ranges. At each split node we then

evaluate on-the-fly a few hundred box features with a brute force search for optimal thresholds over the range of feature responses to greedily find the most discriminative feature/threshold pair. This strategy has proven successful in a number of works using RFs for various tasks.

Deep Learning We also experiment with convolutional neural networks (CNNs) as RCA classifiers. Here, we utilize DeepMedic, a 3D CNN architecture for automatic segmentation [74]. The architecture is computationally efficient as it can handle large image context by using a dual pathway for multi-scale processing. CNNs have been shown to be able to learn highly complex and discriminative data associations between input data and target output. The architecture of the network is defined by the number of layers and the number of activation functions in each layer. In CNNs, each activation function corresponds to a learned convolutional filter, and each filter produces a feature map (FM) by convolving the outputs of the previous layer. Through the sequential application of many convolutions, highly complex features are learned that are then used to produce voxel-wise predictions at the final, fully-connected layer. CNNs are a type of deep learning approach which normally requires large amounts of training data in order to perform well due to the thousands (or millions) of parameters corresponding to the weights of the filters.

To be able to act as a RCA classifier that is trained on a single image, we require a specialised architecture. Here, we reduce the number of FMs in each layer by one third compared to the default setting of DeepMedic. We also cut the feature maps in the last fully connected layers, from 150 to 45. By reducing the feature maps without changing the architecture in terms of number of layers the network preserves its capability to see large image context as the size of the receptive field remains unchanged. With less number of filters, the number of parameters is substantially decreased, which leads to faster computations, but more importantly, reduces overfitting when trained on a single image. Training is performed in a patch-wise manner where the original input image is divided into 3D patches that are then sampled during training using backpropagation and batch normalization. For details about the training procedure and further analysis of DeepMedic we refer to [74].

Atlas-based Label Propagation The third approach we consider is atlas-based label propagation. Label propagation using multiple atlases have been shown to yield state-of-the-art results on many segmentation problems [68]. A common procedure in multi-atlas methods is to use non-rigid registration to align the atlases with the image to be segmented and then perform label fusion strategies to obtain predictions for each image point. Although, multi-atlas methods based on registration are not strictly voxel-wise classifiers as they operate on the whole image during registration, the final stage of label fusion can be considered as a voxel-wise classification step. Here, we make use of an approach that has been originally developed in the context of segmentation of cardiac MRI [7].

For the purpose of RCA, however, there is only a single atlas and thus no label fusion is required. Using single atlas label propagation then boils down to making use of an efficient non-rigid registration technique as the one described in [7]. For RCA, the single atlas then corresponds to the image and its predicted segmentation for which we want to estimate the segmentation quality. We use the same configuration for image registration as in [7] and refer to this work for further details.

3.2.2 Predicting Segmentation Accuracy

For the purpose of assessing the quality of an individual segmentation, we train a RCA classifier f_{I,S_I} on a single image I that has been segmented by any segmentation method, where S_I denotes the predicted segmentation that here acts as pseudo GT during classifier training. Our objective is to estimate the quality of S_I in the absence of GT. To this end, we define the segmentation function $F_{I,S_I}(J) = S_J$ that applies the trained RCA classifier f_{I,S_I} to all voxels (or more precisely to the features extracted at each voxel) of another image J which produces a segmentation S_J. Assuming that for the image J a reference GT segmentation S_J^{GT} is available, we can now compute any segmentation evaluation metric on the pair (S_J, S_J^{GT}). The underlying hypothesis in our RCA framework is that there is a correlation between the values computed on (S_J, S_J^{GT}) and the values one would get for the pair (S_I, S_I^{GT}), where S_I^{GT} is the reference GT of image I which in practice, however, is unavailable.

It is unlikely that this assumption of correlation holds for an arbitrary reference image J. In fact, the RCA classifier f_{I,S_I} is assumed to work best on images that are somewhat similar to I. Therefore, we further assume that a suitable reference database is available that contains multiple segmented images (or atlases) $\mathcal{T} = \{(J_k, S^{\mathrm{GT}}_{J_k})\}_{k=1}^{m}$ that capture the expected variability. Such a database is commonly available in the context of machine learning and multi-atlas based segmentation approaches, but could also be generated specifically for the purpose of RCA. If already available, we can re-use existing training databases that might have been previously used during method development and/or cross-validation and parameter tuning. When testing the RCA classifier on all of the available m reference images, we expect that the RCA classifier performs well on at least some of these, if and only if the predicted segmentation S_I is of good quality. If S_I is of bad quality, we expect the RCA classifier to perform poorly on all reference images. This leads to our definition of a proxy measure for predicting the segmentation accuracy as

$$\bar{\rho}(S_I) = \max_{1 \leq k \leq m} \rho(F_{I,S_I}(J_k), S^{\mathrm{GT}}_{J_k}) \tag{3.1}$$

where ρ is any evaluation metric, such as DSC, assuming higher values correspond to higher quality segmentations[2]. Here, we only look for the maximum value that is found across all reference images, as this seems to be a good indicator of the quality of the segmentation S_I. Other statistics could be considered, such as the average of the top three scores, but we found that the maximum score works best as a proxy. Note, that the mean or median scores are not very useful measures as we do not expect the RCA classifier to work well on the majority of the reference images. Afterall, the RCA classifier does overfit to the single image and will not generalize to perform well on dissimilar images. Nonetheless, as we will demonstrate in the experiments, $\bar{\rho}$ indeed provides accurate estimates for the segmentation quality in a wide range settings.

[2]For metrics where a lower value indicates better quality, such as surface distance, we can simply replace the max with a min operator.

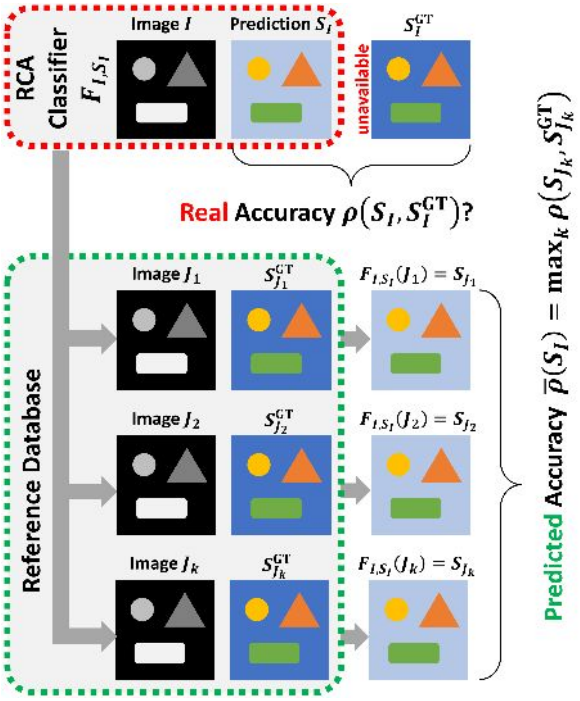


Figure 3.1: Overview of the general framework for reverse classification accuracy. The accuracy of the predicted segmentation S_I of a new image I is estimated via a RCA classifier trained using S_I as pseudo ground truth, and applied to a set of reference images J_k for which ground truth reference segmentations $S_{J_k}^{\mathrm{GT}}$ are available. The best segmentation score computed on the reference images is used as prediction of the real accuracy of segmentation S_I.

3.2.3 Summary

The following provides a summary of the required steps for using RCA in practice within a processing pipeline for automatic image segmentation. Given an image I to be segmented:

1. Run the automated image segmentation method to obtain predicted segmentation S_I.

2. Train a RCA classifier on image I and its predicted segmentation S_I to obtain an image segmenter F_{I,S_I}.

3. Evaluate the RCA classifier on a reference database with images for which GT is available $\mathcal{T} = \{(J_k, S_{J_k}^{\mathrm{GT}})\}_{k=1}^{m}$ to obtain segmentations $\forall k\, F_{,S_I}(J_k) = S_{J_k}$.

4. Compute the segmentation quality of S_I using a proxy measure $\bar{\rho}(S_I)$ according to Eq. (3.1).

Depending on the application, a threshold may be defined on $\bar{\rho}$ to flag up images with poor

segmentation quality that need manual inspection, or to automatically identify high quality segmentations suitable for further analysis.

3.3 Experiments and Results

In order to test the effectiveness of the RCA framework, we explore a comprehensive multi-organ segmentation task on whole-body MRI. In this application, we evaluate the prediction accuracy of RCA in the context of three different state-of-the-art segmentation methods, a Random Forest approach [49], a deep learning approach using 3D CNNs [74], and a probabilistic multi-atlas label propagation approach [7]. The dataset used to validate our framework is from our MALIBO study. We collected whole-body, multi-sequence MRI (T1w Dixon and T2w images) of 35 healthy volunteers. Detailed manual segmentations of 15 anatomical structures, including abdominal organs (heart, left/right lung, liver, adrenal gland, gall bladder, left/right kidney, spleen, pancreas, bladder) and bones (spine, left/right clavicle, pelvis) have been generated by clinical experts as part of the study. These manual segmentations will serve as GT in the quantitative evaluation.

3.3.1 Experimental Setting

We use 3-fold cross validation to automatically segment all 525 structures (15 organs $\times$ 35 subjects) with each of the three different segmentation methods, namely Random Forests, CNNs, and Multi-Atlas. In each fold, we use the RCA framework with three different methods for realizing the RCA classifier, namely Atlas Forests, constrained CNNs, and Single-Atlas, as described above. Using the RCA classifiers that are trained on each image for which we want to assess segmentation quality, we obtain segmentations on all reference images which are then compared to their manual reference GT. Since the GT is available for all 35 cases, we can compare the predicted versus the real segmentation accuracy for all cases and all organs under various settings with nine different combinations of segmentation methods and RCA classifiers.

3.3.2 Quantifying Prediction Accuracy using Different Reverse Classifiers

The Dice's similarity coefficient is the most widely used measure for evaluating segmentation performance, and in our main results we focus on evaluating how well DSC can be predicted using our RCA framework. In order to quantify prediction accuracy, we consider three different measures, namely the Pearson's correlation coefficients between predicted and real DSC, the mean absolute error (MAE), and a classification accuracy. Arguably, the most important measure for direct evaluation of how well RCA works is the MAE, as it directly tells us how close the predicted DSC is to the real one. Correlation is interesting, as it tells us something about the relation between predicted and real scores. We expect high correlation in order for RCA to be useful, but we might not always have an identity relation, as there could be a bias in the predictions. For example, if the predicted score is consistently lower than the real score, this can still be useful in practice, and will be indicated by high correlation but might not yield low MAEs. In such a case, a calibration might be considered as we will discuss later on.

We also explore whether the predictions can be used to categorize segmentations according to their quality. We argue that for many clinical applications it is already of great value to be able to discriminate between good, bad, and possibly medium quality segmentations and that the absolute segmentation scores are of less importance. For proof-of-principle, we consider a three-category classification by grouping segmentations within DSC ranges $[0.0, 0.6)$ for bad, $[0.6, 0.8)$ for medium, and $[0.8, 1.0]$ for good cases. Note, that those ranges are somewhat arbitrary, in particular, as the quality of absolute DSC values is highly depending on the structure of interest. So in practice, those ranges would need to be adjusted specifically to the application at hand.

3.3.3 Results for Predicting Dice's Similarity Coefficients

Our main results are summarized in Table 3.1 where we report the quantitative analysis of the predicted accuracy for nine different settings consisting of three different segmentation methods and three different ways of realizing the RCA classifier. In Figure 3.2 we provide the scatter

Table 3.1: Predicting DSC for different segmentation methods using different RCA classifiers. Best results are obtained with the Single-Atlas approach yielding the lowest error between real and predicted DSC scores and high classification accuracy.

		Correlation		MAE		Accuracy 3-Categories	
Segmentation Method	RCA Classifier	All	No Zeros	All	No Zeros	All	No Zeros
Random Forests	Atlas Forests	0.881	0.867	0.120	0.130	0.783	0.776
CNNs	Atlas Forests	0.828	0.630	0.166	0.245	0.623	0.500
Multi-Atlas	Atlas Forests	0.863	0.877	0.168	0.177	0.749	0.726
Random Forests	Constrained CNNs	0.721	0.718	0.252	0.271	0.653	0.631
CNNs	Constrained CNNs	0.756	0.662	0.225	0.292	0.592	0.472
Multi-Atlas	Constrained CNNs	0.773	0.686	0.209	0.237	0.693	0.642
Random Forests	Single-Atlas	0.955	0.946	0.051	0.052	0.888	0.880
CNNs	Single-Atlas	0.973	0.892	0.052	0.065	0.811	0.756
Multi-Atlas	Single-Atlas	0.962	0.947	0.067	0.072	0.822	0.798

plots of real versus predicted DSC for all nine settings with 525 data points each.

Overall, we observe high correlation between predicted and real DSC for both Atlas Forests and Single-Atlas when used as RCA classifiers, with the Single-Atlas showing correlations above 0.95 for all three segmentation methods. The Single-Atlas approach also yields the lowest MAEs between 0.05 and 0.07, and good 3-category classification accuracies between 81% and 89%. This is visually confirmed by the scatter plots in the right column of Figure 3.2 which show good linear relation close to the diagonal between predicted and real scores for most structures in the case where Random Forests or Multi-Atlas are used as the original segmentation method. When using Atlas Forests for RCA, we still observe good correlation but the relationship between predicted and real scores is off-diagonal with larger spread towards lower quality segmentation. The correlation is still good and above 0.82, MAEs are between 0.12 and 0.17 with classification accuracy going down to 0.62%, 0.75% and 0.78% depending on the original segmentation method. For the case of the constrained CNNs, we observe that the prediction quality is lowest confirmed by the scatter plots and all quantitative measures, with correlations below 0.78 and MAEs above 0.2. The constrained CNNs seem to only work for predicting segmentation accuracy in case of major organs such as liver, lungs, and the spine but clearly struggle with smaller structures leading to many zero predictions even when the real DSC is rather high. This is most likely caused by the difficulty of training the CNNs with single images and small structures which does not provide sufficient amounts of training data.

Figure 3.3 shows an example for predicting the accuracy of a liver segmentation. Next to a

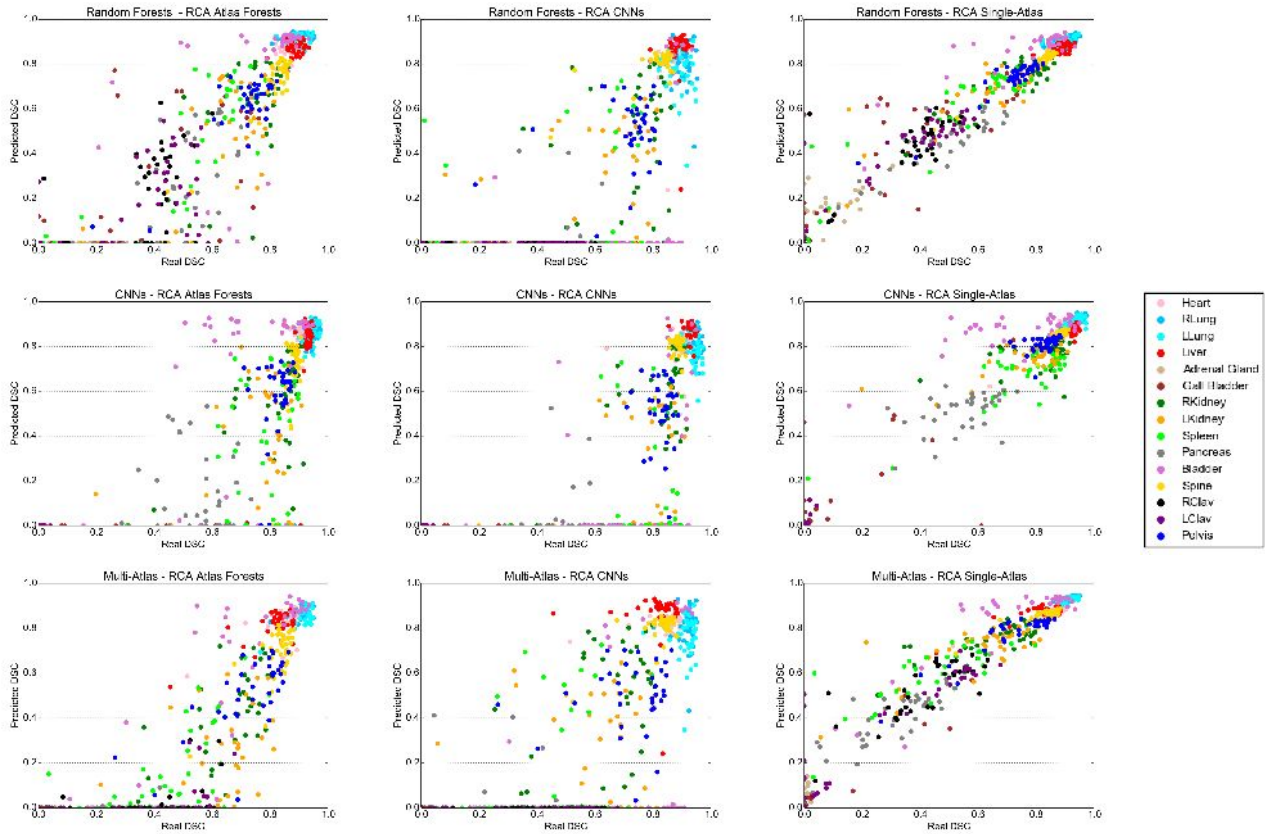


Figure 3.2: Scatter plots of predicted and real DSC of multiple structures for three different segmentation methods (rows) using three different RCA classifiers (columns). High correlation and low prediction errors are obtained when employing the single-atlas label propagation as RCA classifier (right column). There is also good correlation with predictions in case of Atlas Forests (left) with larger spread towards lower quality segmentations. The constrained CNNs (middle column) are less suitable for RCA which is likely due to the difficulty of training on single images. Both, Atlas Forests and constrained CNNs work best for larger organs such as liver, lungs, and pelvis while leading to many zero predictions for smaller structures such as adrenal gland and clavicles. The single-atlas label propagation makes accurate predictions of segmentation quality across all 15 anatomical structures. A summary of the plots is given in Tab. 3.1.

slice from a T2w MRI volume we show the GT manual segmentation together with the result from a Random Forest. Underneath, we show the 24 segmentations obtained on the reference database when using the Single-Atlas RCA approach. The bar plot in the same figure shows the variation of the 24 DSC scores. Similarly, the bar plots in Figure 3.4 of two more examples illustrate the distribution of DSC scores when predicting a good quality segmentation on the left, and a poor quality segmentation on the right. The three examples support the hypothesis that selecting the maximum score across the reference database according to Eq. (3.1) is a good proxy for predicting segmentation quality.

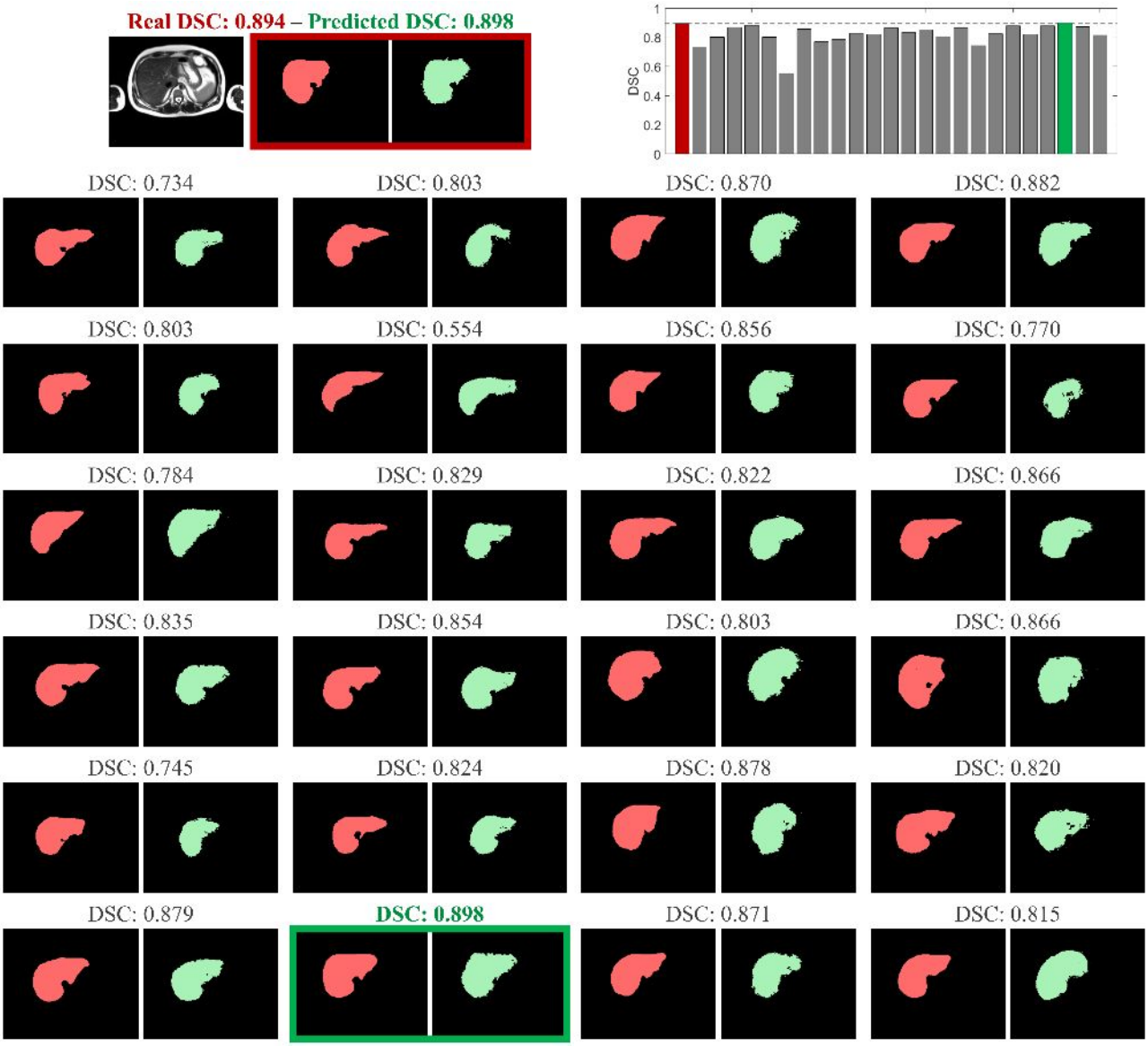


Figure 3.3: Visual examples for predicting the segmentation quality of the liver for a new image shown on the top. Its GT segmentation (red colored) is unknown in practice, and we want to estimate the quality of the predicted, automatic segmentation shown on the most right (green colored). By taking the predicted segmentation as pseudo GT and training a RCA classifier we can obtain segmentations on a reference database with 24 images with available GT. The bar plot shows the real DSC in red and the different DSC values obtained for the reference images shown below. The green bar corresponds to the maximum DSC, and is selected as predicted DSC for the new image according to Eq. (3.1) which matches well the real DSC.

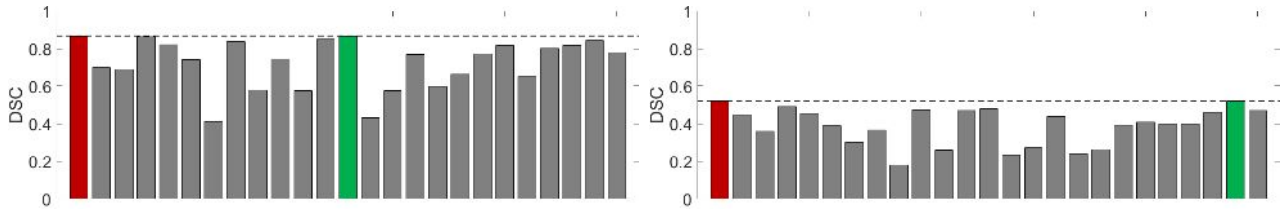


Figure 3.4: The bar plots show two examples for predicting the real DSC (red) in case of a good quality (left) and bad quality segmentation (right) using a database with 24 reference images with available GT. The predicted DSC (green bar) selected according to Eq. (3.1) matches well the real DSC.

Some of the original segmentation methods have problems segmenting structures such as the adrenal gland and clavicles. The CNNs, in particular, failed to segment adrenal glands in most cases. Because the real DSC for these is zero with no voxels labelled in the segmentation map, the RCA predictions are always correct as there are no labels for the RCA classifier for this structure. In order to investigate the effect of those zero predictions on the quantitative results, we also report in Table 3.1 under the columns 'No Zeros' the correlations, MAEs and classification accuracies when structures with a real DSC of zero are excluded. We observe that the zero predictions have mostly an impact on CNNs, either employed as original segmentation method or as RCA classifier. For Atlas Forests and Single-Atlas the effect on the accuracies is very little, confirming that those both are well suited within the RCA framework, independent of the original segmentation method.

An example on how our RCA framework works in different organ segmentation and methods is shown in Figure 3.5. Using the predicted DSC by RCA, we can rank the performance of segmentation methods in a particular organ. RCA can rank the DSC for good segmentation quality (see lung and liver segmentation in Figure 3.5-top and middle), as well as low segmentation quality, given in Figure 3.5-bottom for pancreas segmentation. For lung and liver segmentation, predicted DSCs by RCA shows that CNN segmentation has the best quality, followed by forest segmentation and multi-atlas segmentation, which are also shown by the real DSCs.

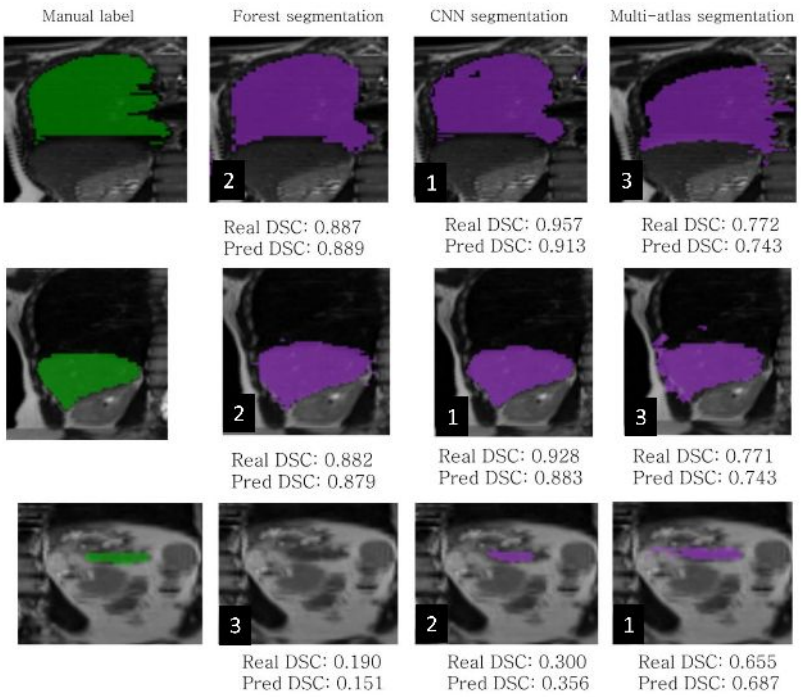


Figure 3.5: Examples on predicting DSC from different segmentation methods on organ segmentation. RCA gives a good agreement to rank the best segmentation methods and to predict the DSCs in right lung segmentation (top) and liver segmentation (middle). Bottom: Prediction of low DSCs on pancreas segmentation also gives a good agreement to rank the best methods.

3.3.4 Predicting Degraded Segmentation Accuracies

In clinical routine it is of great importance to be able to detect when an automated method fails. We conducted a dedicated experiment to investigate how well RCA can predict segmentation failure. From the scatter plots in Figure 3.2 we can see that all three segmentation methods perform reasonably well on most major organs with no failure cases among structures such as liver, heart, and lungs. In order to further demonstrate that RCA can predict failure cases in these structures, we utilize degraded Random Forests by limiting the tree depth at test time to 8. This leads to much worse segmentation results for most structures which is confirmed in the corresponding scatter plots shown in Figure 3.6.

Again, we evaluate the performance of the three different RCA classifiers, Atlas Forests, constrained CNNs and Single-Atlas. The results are summarized in Table 3.2. The constrained CNNs are again suffering from many zero predictions and less suitable for making accurate predictions. Atlas Forests and Single-Atlas, however, result in high correlations, low MAEs

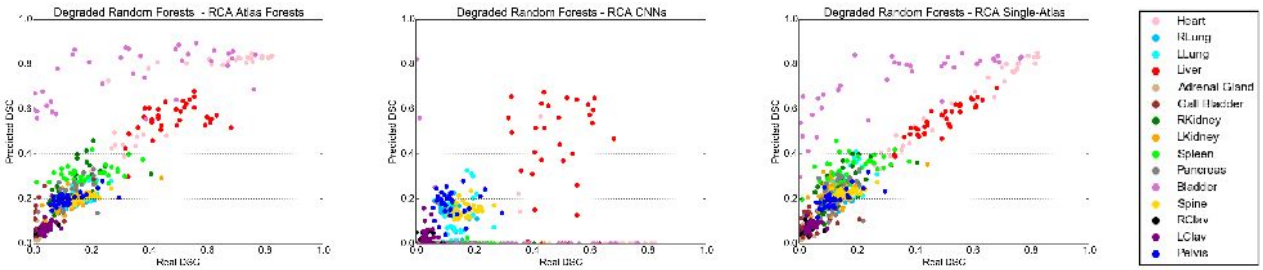


Figure 3.6: Scatter plots for the experiment for detecting segmentation failure when using degraded Random Forests with limited depth as the segmentation method. Atlas Forests (left) and single-atlas label propagation (right) make highly accurate predictions in the low DSC ranges and thus, are able to correctly detect such failed segmentations, with the exception of the bladder. Constrained CNNs are again less suitable for RCA with many zero predictions.

Table 3.2: Detecting segmentation failure

RCA Classifier	Correlation	MAE	Accuracy 3-Cat.
Atlas Forests	0.853	0.096	0.884
Constrained CNNs	0.237	0.139	0.924
Single-Atlas	0.875	0.097	0.928

and very good classification accuracies. Low real DSC scores are correctly predicted and failed segmentations are identified. The only exception here is the bladder. This might be explained by the unique appearance of the bladder in the multi-spectral MRI with hyper-intensities in the T2w image, and its largely varying shape between subjects. It appears that even a badly segmented bladder can be sufficient for the RCA classifier to learn its appearance and segment the bladder well on at least one of the reference images. Overall, the experiment suggests that RCA with Atlas Forests and Single-Atlas can be employed in automatic quality control, for example, in large-scale studies where it is important to be able to detect failed segmentations which should be excluded from subsequent analyses.

3.3.5 Results for Predicting Different Segmentation Metrics

We further explore the ability to predict other evaluation metrics than DSC. We consider the following metrics: Jaccard index (JI), precision (PR), recall (RE), average surface distance (ASD), Hausdorff distance (HD) and relative volume difference (RVD). In the following is description of each quantitative metric we use, between the automatic segmentation (AS) and the ground truth (GT).

Dice similarity coefficient

Most of our experiment use DSC as it is the most widely used metric to evaluate a segmentation result. DSC is obtain by Eq.3.2

$$DSC = \frac{2|AS \cap GT|}{|AS| + |GT|} \tag{3.2}$$

Jaccard index (JI)

Another commonly used metric is Jaccard, which is the number of common voxels of the segmentation results and ground truth region over their union:

$$Jaccard = \frac{|AS \cap GT|}{|AS \cup GT|} \tag{3.3}$$

Precision and Recall

The similarities in the volumes of segmented regions and the ground truth can be reflected by the precision and recall. Precision and recall together can form the F1-measure, which is equivalent to DSC [132].

$$Precision = \frac{|AS \cap GT|}{|AS|} \tag{3.4}$$

$$Recall = \frac{|AS \cap GT|}{|GT|} \tag{3.5}$$

Overall, we achieve a strong correlation between the predicted and the real overlap based metrics, low MAE, and high accuracy in all categories (see Table 3.3). Since we can compute the Jaccard from the DSC, it is not surprising that the predicted Jaccard performs as good as in DSC. Meanwhile, precision performs lower than recall in terms of correlation and accuracy in categorization. This is due to the different portion of voxels measured in each metrics [132]; under-segmentation error is not reflected in the precision, while over-segmentation error is not measured in the recall [153].

- Surface distance based methods

Hausdorff Distance (HD)

The Hausdorff distance between two sets of points AS and GT defines the maximum distance (h) from a point in $AS \rightarrow GT$ and $GT \rightarrow AS$. [5]

$$HD(AS, GT) = max(h(AS, GT), h(GT, AS)) \tag{3.6}$$

Average Surface Distance (ASD)

Average Surface Distance gives the distance error by computing the the average of all distances on the boundary of AS to GT [153]

$$HD(AS, GT) = mean(h(AS, GT), h(GT, AS)) \tag{3.7}$$

The distance errors obtained from the surface distance based methods are unbounded, which depend on the size of the segmentation and organ. The results of the predicted HD and ASD shown in Table 3.3 give a weaker correlation than the overlap based metrics, but the MAE is considerably fair. Different ranges in HD and ASD also make the different splits in categorizations. The accuracy of category classification in HD is better than ASD and precision, while ASD has very low accuracy.

- Volume based methods

Relative Volume Difference (RVD)

Finally, we also look at the volume, as the number of labelled voxels multiplied by the voxel spacing. We calculate the RVD as the ratio of the absolute difference between AS and GT, to the GT volume [5].

$$RVD = \frac{|V_{AS} - V_{GT}|}{V_{GT}} \tag{3.8}$$

The values of RVD are also unbounded, depend on the ratio of volume under or over-segmentation. Therefore, for this case, we also set the range that the segmentation with RVD above 1 is considered as failed. However, the predicted RVD have a weak correlation and low accuracy on

all categories (see Table 3.3). For this experiment, we use Random Forests as segmentation method, and Atlas Forests for RCA. The results are summarized in Table 3.3.

Good correlation is obtained between predicted and real overlap based scores, with low MAEs, and high accuracies. Since Jaccard is directly related to DSC[3], it is expected that the predictions are of similar quality. Prediction accuracy for precision is lower than for recall. The two metrics capture different parts of segmentation error; under-segmentation is not reflected in precision, while over-segmentation is not captured in recall[4].

Distance based errors are unbounded, so we define thresholds for HD and ASD, and errors above are clipped to the threshold value, which is set to 150mm for HD, and 10mm for ASD. This also allows us to define ranges for the error categorization. For HD, we use the ranges $[0, 10]$, $(10, 60]$, and $(60, 150]$ for good, medium and bad quality segmentations. For ASD we divide the range into $[0, 2]$ for good, $(2, 5]$ for medium, and $(5, 10]$ for bad segmentation quality. Compared to overlap based metrics, the RCA predictions for HD and ASD are not convincing with low correlation, high MAE, and low classification accuracy.

RVD is the ratio of the absolute difference between the reference and predicted segmentation volume and the reference volume. Perfect segmentation will result in a value of zero. As RVD is also unbounded, we use a threshold of one to indicate maximum error. The predictions for RVD are good, with high classification accuracy of 0.68%, similar to the overlap based scores. In conclusion, it seems RCA works very well for overlap based measures and for RVD to some extent, while distance based metrics cannot be accurately predicted with the current setting and would require further investigation.

3.4 Discussion and Conclusion

The experimental validation of the RCA framework has shown that it is indeed possible to accurately predict the quality of segmentation in the absence of GT, with some limitations.

[3] $JI = DSC/(2 - DSC)$.
[4] $DSC = 2 \cdot PR \cdot RE/(PR + RE)$.

Table 3.3: Predicting different segmentation metrics

Metric	Correlation	MAE	Accuracy 3-Cat.
DSC [0 - 1]	0.881	0.120	0.783
JI [0 - 1]	0.899	0.110	0.749
PR [0 - 1]	0.572	0.242	0.638
RE [0 - 1]	0.833	0.170	0.663
HD [0-150]	0.189	41.38	0.387
ASD [0-10]	0.022	4.120	0.344
RVD [0-1]	0.474	0.051	0.678

Table 3.4: Comparison of predicting DSC with and without calibration via regression.

		Correlation		MAE		Accuracy 3-Categories	
Segmentation Method	RCA Classifier	Direct	Calibrated	Direct	Calibrated	Direct	Calibrated
Random Forests	Atlas Forests	0.881	0.833	0.120	0.105	0.783	0.809
CNNs	Atlas Forests	0.828	0.929	0.166	0.079	0.623	0.783
Multi-Atlas	Atlas Forests	0.863	0.939	0.168	0.065	0.749	0.745
Random Forests	Constrained CNNs	0.721	0.826	0.252	0.104	0.653	0.768
CNNs	Constrained CNNs	0.756	0.961	0.225	0.049	0.592	0.830
Multi-Atlas	Constrained CNNs	0.773	0.874	0.209	0.100	0.693	0.787
Random Forests	Single-Atlas	0.955	0.872	0.051	0.089	0.888	0.815
CNNs	Single-Atlas	0.973	0.967	0.052	0.048	0.811	0.811
Multi-Atlas	Single-Atlas	0.962	0.918	0.067	0.080	0.822	0.825

We have explored different methods for realising the RCA classifier and could demonstrate that Atlas Forests and in particular, Single-Atlas label propagation yield accurate predictions for different segmentation methods. As the RCA framework is generic, other methods can be considered and it might be necessary to select the most appropriate one for the application at hand. We have also experimented with a constrained CNN trained on single images, which only works well for major organs such as liver, lungs and spine. There might be other more appropriate architectures for the purpose of RCA, which will be explored as part of future work.

An appealing property of the proposed framework is that unlike the supervised methods in [45] and [76] no training data is required that captures examples of good and bad segmentations. Instead, in RCA we simply rely on the availability of a reference database with available GT segmentations. The drawback, however, is that we assume a linear relationship between predicted and real scores which should be close to an identity mapping, something we only found in the case of using Single-Atlas label propagation (see right column of Figure 3.2). In the case of off-diagonal correlation, as for example found for Atlas Forests, an extension to RCA could be considered where the predictions are calibrated. This, however, requires training data

from which a regression function could be learned, similar to [76]. In order to demonstrate the potential of such an approach, we perform a simple experiment on the data that we used for conducting the main evaluation.

We have tried several regression algorithms with cross-validation training set, such as linear regression, random forest (RF) regression [16], gradient boosted regression tree (GBRT)[44], and SVM regression (SVR) [128] with different kernels. However, RF regression gave the highest correlation. Compared to [76] that uses segmentation features as the input for regression, our method directly uses the performance proxy from RCA. After obtaining all predicted DSC scores, we run a leave-one-subject-out validation where in each fold we use Random Forest regression to calibrate the predictions. The results are summarized in Tab. 3.4 where we compare the quantitative measures before and after calibration. Both the MAEs and classification accuracies improve significantly for the case of Atlas Forests and constrained CNNs. For Single-Atlas, however, the results remain similar due to the already close to identity relationship between predicted and real scores before calibration. Calibration, however, comes with a risk of overfitting as the method will learn the relationship on the available training data but might not generalize to new data.

In our experiments we have found that best predictions are obtained for overlap based measures such as DSC and Jaccard Index. Whether those measures are sufficient to fully capture segmentation quality is debatable. Still, DSC is the most widely considered measure and being able to accurately predict DSC in the absence of GT has high practical value. Besides being useful for clinical applications where the goal is to identify failed segmentations after deployment of a segmentation method, we see an important application of RCA in large-scale imaging studies and analyses. In settings where thousands of images are automatically processed for the purpose of deriving population statistics, it is not feasible to employ manual quality control with visually inspection of the segmentation results. Here, RCA can be an effective tool to automatically extract the subset of high quality segmentations which can be used for subsequent analysis. We are currently exploring this in the context of population imaging on the UK Biobank imaging data where image data of more than 10,000 subjects is available which will be subsequently increased to 100,000 over the next couple of years. The UK Biobank data

will enable the discovery of imaging biomarkers that correlate with non-imaging information such as lifestyle, demographics, and medical records. In the context of such large scale analysis, automatic quality control is a necessity and we believe the RCA framework makes an important contribution in this emerging area of biomedical research. In the next chapter, we will further explore the use of RCA for other image analysis and segmentation task.

Chapter 4

Learning between Different Domains

This chapter is based on:

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker. Domain Adaptation for MRI Organ Segmentation using Reverse Classification Accuracy. *1st International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

In this chapter, we explore a useful application of reverse classification accuracy (RCA) framework introduced in Chapter 3. The concept of RCA is applied to the domain adaptation problem where normally there is no ground truth available in the target domain. The variations in multi-center data in medical imaging studies have brought the necessity of domain adaptation. Despite the advancement of machine learning in automatic segmentation, performance often degrades when algorithms are applied on new data acquired from different scanners or sequences than the training data. Manual annotation is costly and time consuming if it has to be carried out for every new target domain. In this work, we investigate automatic selection of suitable subjects to be annotated for supervised domain adaptation using the concept of RCA.

RCA predicts the performance of a trained model on data from the new domain. Then, different strategies of selecting subjects to be included in the adaptation via transfer learning are evaluated. We perform experiments on a two-center MR database for an organ segmentation

task. We show that subject selection via RCA can reduce the burden of annotation of new data for the target domain.

4.1 Introduction

Machine learning has led to significant advances in medical imaging, particularly with big improvements in medical image segmentation. Performance, however, depends on the availability of sufficient amounts of labeled samples for supervised learning, and also whether the test data is coming from the same domain as the training data. In clinical practice, the source domain (S) on which the classifier is trained might be different from the target domain (T) with clinical data. The images from these domains are samples from different appearance distributions. The mismatch of distributions is caused by various factors such as the use of different scanners, types of sequences, or biases in patient cohorts - often causing a trained algorithm to perform poorly on new data. In a scenario where the tasks are the same, but the source and target domains are different, domain adaptation is usually performed to address the domain disparity problem [108].

Domain adaptation can be categorized into three settings, supervised, semi-supervised, and unsupervised [120, 72]. Our work focuses on supervised domain adaptation methods, which uses labeled data from the target domain. In the context of convolutional neural networks (CNNs), supervised domain adaptation can be approached by training from scratch or fine-tuning a network pre-trained on the source domain [124].

An approached called DALSA [55] explored a supervised domain adaptation by introducing a weighting scheme in Random Forests and SVMs for segmentation. This approach preserves the segmentation quality even though only sparse annotated data is used. Another approach tackles the segmentation difficulty of images from different scanners and imaging protocols by weighting different SVM-based classifiers for transfer learning [144]. More recent work has tried to explore transfer learning in CNNs in medical imaging [126, 133], which confirmed the potential of fine-tuned and fully trained CNNs.

However, the importance of subject selection in transfer learning has not been widely studied yet. Most works [4, 120, 167] attempt to integrate active learning with domain adaptation. Intuitively, active learning is chosen since it develops a criterion to determine the "value" of a candidate for annotation.

In active domain adaptation, the classifier selects among the limited labels on target data [120] by combining hybrid oracles. However, obtaining the oracles is costly. Instead of selecting instances to be added to the training set, [4] selects a bag of instances by self-training. But, this approach is more effective when the source and target domains differ substantially. An appealing work is a recent application of active learning in domain adaptation for biomedical images [167]. Active learning chooses the candidates with higher entropy and higher diversity, which are expected to improve the current performance. In [167], a pre-trained CNN is further fine-tuned continuously by incorporating newly annotated samples in each iteration to enhance the performance incrementally. Although they can reduce the annotation cost, the iterative scheme can be time consuming, especially if applied to volumetric data. Previous work on subject selection with active learning [4, 120, 167] requires iterations and only deals with classification of 2D images.

Instead of using iterative active learning, we propose a framework to automatically select the "most valuable" samples from the unlabeled target domain to be annotated - using reverse classification accuracy (RCA) [140]. This chapter will explain one of the useful applications of RCA framework, which is detailed in previous chapter.

4.2 Methodology

4.2.1 Datasets

Source domain (S): The dataset is obtained from our MALIBO study (MAchine Learning In whole Body Oncology) and includes abdominal T1-weighted MR Dixon images of 35 healthy subjects. We consider this dataset as the source domain used to train the initial classifier in

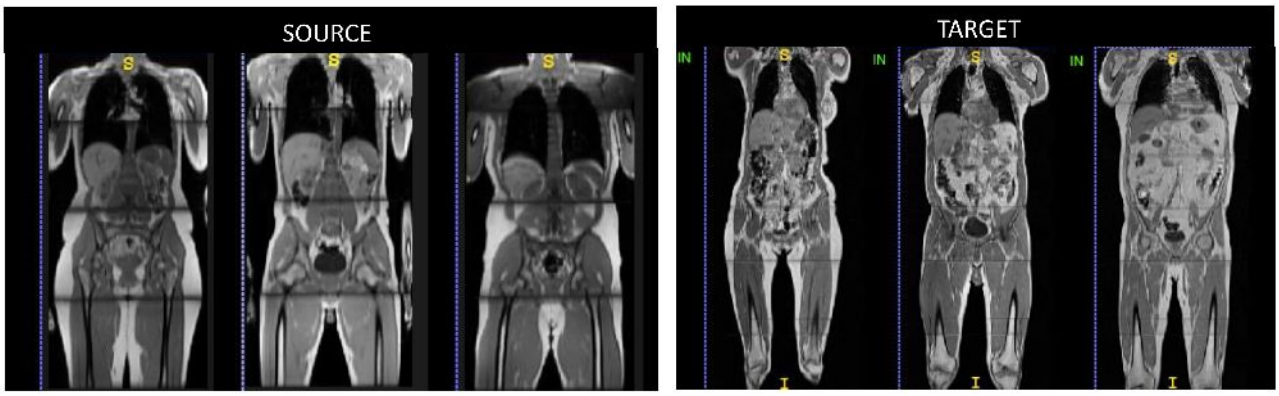


Figure 4.1: Examples of whole-body scans from source (MALIBO) and target (UK Biobank) database.

a supervised manner as manual organ annotations are available for all subjects. The images have size of $(256 \times 208 \times 202)$ and resolution $(1.64 \times 1.64 \times 5)$ mm.

Target domain (T): Data for the target domain is obtained from the UK Biobank (Application 12579). We use 45 subjects with manually annotated T1-weighted Dixon MR images which have been acquired with a similar protocol as the MALIBO data. The main obvious differences are the image size $(224 \times 168 \times 366)$ and resolution $(2.23 \times 2.23 \times 3)$ mm.

Source and target dataset are acquired at different centers. UK Biobank data is acquired with a Siemens 1.5T MAGNETOM Aera scanner while in MALIBO a Siemens 1.5T MAGNETOM Avanto is used. For our study, we use the T1-weighted in-phase images from the Dixon protocol. We focus on organ segmentation within whole-body scans. As a pre-processing, image intensities in both datasets are normalized to zero-mean and unit-variance. Images are resampled to the same size and physical resolution. Visual examples from the source and target database are depicted in Figure 4.1. Despite the similarity of the scanning protocol and same scanner manufacturer, the drop in segmentation accuracy when applying a model trained on MALIBO and tested on UK Biobank data is striking, as we will show in our experiments. The images seem to encode a significant bias in their appearance which is not obvious upon visual inspection.

4.2.2 Domain Adaptation using Reverse Classification Accuracy (DARCA)

As we have explained in the previous chapter, RCA can accurately predict the quality of a segmentation on a case-by-case basis without the need for labeled testing set. Here we apply RCA with the single-atlas registration classifier. The test image together with its predicted segmentation is registered to a set of reference images such that the predicted segmentation can be quantitatively compared to the manual segmentations of the reference images by computing Dice similarity coefficients (DSC). It is expected that the maximum DSC score over all reference images correlates well with the real DSC. For more details on the RCA framework, see Chapter 3.

We address the question whether RCA can be employed to select few subjects to reduce the cost of annotation in supervised domain adaptation. RCA acts as a selector for picking up subjects with high and low confidence in segmentation accuracy. Our hypothesis is that transfer learning with specific n-subject selection is better than picking-up random subjects from the target domain, and thus fewer manually labelled subjects are needed from the target domain. In the following, we call this framework as 'domain adaption using RCA', or DARCA.

To answer the usefulness of DARCA, in the experiments we demonstrate the effective use of RCA as a selector for n-subjects in target domain to be incorporated into the training dataset, we compare different strategies for supervised domain adaptation with the RCA selection, and we study how the training size and the combination of target samples affects segmentation performance.

4.2.3 Supervised Domain Adaptation

In our experiments, we employ DeepMedic[1] [74] as the base network for 3D organ segmentation. We use the default 11-layers deep, multi-scale, parallel convolutional pathways architecture.

[1]Source code available at `https://github.com/Kamnitsask/deepmedic`

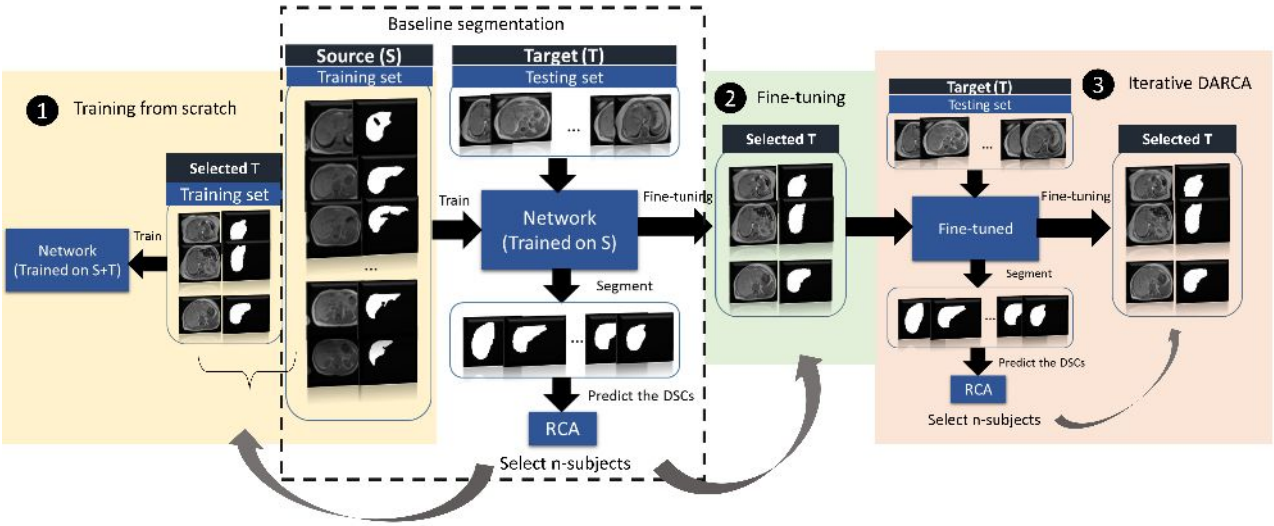


Figure 4.2: Overview of the strategies we tried for n-subject selection in DARCA: 1. Training from scratch. 2. Fine-tuning. 3. Iterative DARCA

The main approaches of supervised domain adaptation with CNNs are either training from scratch or fine-tuning [133]. Using RCA as a selector, n-subjects from the target domain are added to the training data. With the new training data containing source and n-target (S+T) subjects, we train the network from scratch. Meanwhile, we can transfer the parameters from a pre-trained network and fine-tune on another database. To build the model, we fine-tune the pre-trained network (from S-only training), with the n-target subjects selected by RCA. Based on the results in [50], fine-tuning the last layer achieves the best performance compared to using more convolutional layers for fine-tuning. We fine-tuned only the last layer of the pre-trained networks and used the same optimization but with fewer epochs. Based on our experiments fine-tuning all the layers is shown to have lower accuracy and more training time is needed.

4.3 Experiments and Results

We present results for using different strategies to investigate the effect of RCA-based subject selection for domain adaptation, as shown in Figure 4.2. We use 3-fold cross-validation with the same random splits in all experiments. As the baseline, we trained the network with all S data and tested it to segment the T images. We predict the DSCs of all target segmentations using RCA. After we sort their DSCs (lowest to highest confidence), we select n-subjects from

T domain to be included with their corresponding manual annotations in the training set mimicking an active learning approach.

4.3.1 Training from Scratch

We set our baseline as the segmentation of T with an S-only trained network, whereas the upper-bound is the segmentation of T with T-only trained network on liver segmentation, as shown in Table 4.1.

Figure 4.2-1 shows that RCA selects n-subjects from T domain, to be manually labeled and incorporated into training dataset. For this experiment, we compare best-/worst-5 subjects selected by RCA and the real best-/worst-5 (real DSCs from the target ground-truth). Besides, we also run random-5 subject selection (repeated with three different random combination and taking the average) to be trained from scratch.

In the training-from-scratch strategy, we train S+T data simultaneously with the same optimization, update-rule, number of epochs, loss function, and regularization techniques as in the baseline. From Table 4.2 column one, we can see that picking up 5 subjects from target domain has already improved the accuracy, compared to the baseline. Table 4.2 shows that the segmentation accuracy using the best-5 or worst-5 subject selection outperformed the random selection. Moreover, RCA selection gives a relatively similar segmentation accuracy to the selection using real DSC.

As we increase the number of annotated target images to be incorporated in training from scratch, the accuracy does improve, as shown in Figure 4.3. From Table 4.2, T segmentation with lowest confidence also gives a good contribution to the current CNN, therefore we combine 5 target subjects with highest confidence and 5 subjects with lowest confidence into the training. The result shows that this combination achieves similar accuracy to the one that incorporated all annotated T subjects. Hence, we reduce the "less valuable" samples for training by RCA selection. Unfortunately, training from scratch requires more time since the networks must learn from the beginning.

Table 4.1: Baseline and upper-bound accuracies. The upper-bound gives the highest performance when the network is trained and tested on one domain (target dataset). The performance drops significantly when training and testing data are from different domains (baseline).

Training	**DSC (mean (stdv))**
Train on S (baseline)	0.639 (0.149)
Train on T (upper-bound)	0.873 (0.046)

4.3.2 Fine-tuning a Pre-trained Network

In Figure 4.2-2, we fine-tune the pre-trained network (from S), with n-subjects selected by RCA. Fine-tuning requires less time than training from scratch. Here, we fine-tune with three different selections (random, real DSC, and RCA) at different set size (2, 5, 10, 15, and all of the T data). Table 4.2 shows the results when we fine-tune using all of the T data which is very similar (DSC: 0.830) to when we train from scratch (DSC: 0.831).

However, RCA seems worse to predict the real segmentation accuracy of fine-tuning. There is a gap between best-/worst-5 real selection and best-/worst-5 RCA accuracy (Table 4.2). One of the reasons could be due to the under-estimation of RCA prediction on the baseline segmentation accuracies on T, with 0.88 correlation and 0.15 MAE. As explained in the experiments of multiple-organs segmentation [140], the RCA prediction for organs with the real DSCs between (0.6 - 0.8) is not as accurate as in organs with real DSCs above 0.8. In our case, the average real DSCs of baseline liver segmentation is 0.639, but RCA under-estimated the mean of predicted DSCs to be 0.497.

As we did in Section 4.3.1, we also combined best-5 and worst-5 in T domain to be annotated and incorporated in fine-tuning. This 10 combined subjects give much better accuracy than when we choose only the best-10 subjects. As shown in (Figure 4.3), fine-tuning with 10 subjects (real selection) gave the highest accuracy (among 2, 5, 15, 20 subjects). Similarly, picking up the worst sample in increasing number, results in lower accuracy than the 'best 5 and worst 5'.

Best-5 and worst-5 annotated samples with real selection shows the best results (DSC: 0.842), and RCA selection also gives similar results (DSC: 0.835), compared to best 10, worst 10, and random 10 by real selection. Additionally, this best-5 and worst-5 combination (real and RCA

selection) performs better than when we use all of the annotated subjects from T (DSC: 0.830). Hence, we cut the cost of annotation by 67%, using only 10 selected subjects instead of 30 subjects and achieve a higher accuracy.

Paired two-sided t-tests are performed for significance analysis against best-5 worst-5 selected by RCA, where the p-values indicate that that selection performed significantly better than the comparison methods (baseline, random 10, worst-10 real selection, and best-10 real selection), see Table 4.3. All of the p-values in the Table 4.3 are less than 0.05 ($p < 0.05$). This means that real difference between Best-5 and Worst-5 performance is significantly different from random 10, worst-10, and best-10 sample selection.

Pseudo-labels for Fine-tuning

If the target domain has no or limited labeled information, adapting CNN to the target domain via fine-tuning is impossible or prone to over-fitting. Therefore, we also investigated the use of pseudo-labels in semi-supervised way. In Section 4.3.2, we incorporated the n-subjects by fine-tuning with their ground-truth. What if, instead of using the real annotations, we use the predicted labels as pseudo ground-truth, which are the baseline segmentation results - for training. In the previous work by [87], pseudo-labels are used for semi-supervised learning with pre-trained and fine-tuning scheme. Pseudo-labels are defined as labels that have maximum predicted probability and seen as equivalent to entropy regularization, which encourages low density separation between classes. In [87], they train using labeled and unlabeled data simultaneously with pseudo-labels (from unlabeled data) to recalculated every weights update for the same loss function.

However, from Figure 4.3, it is clear that using pseudo-labels cannot improve the segmentation performance on the target domain. Fine-tuning with all of the pseudo-labelled subjects in T gives the worst result amongst all. The noisy labels negatively impact the segmentation performance. Interestingly, training from scratch with labeled S and pseudo-labeled T data makes a little improvement (0.652) from the baseline segmentation (0.638). Meanwhile, fine-tuning with all of the pseudo-labeled T data gives the worst results amongst all. The more noisy

data (pseudo-labels) are introduced to the networks, the poorer the segmentation performance. We also investigated the effects of using pseudo-labels in fine-tuning with different size of top-n subject selections. However, the trend of accuracies is unclear in top-n RCA selection, where it increases in top 5 and top 15 selection.

Hence, training using pseudo-labels seems not suitable for domain adaptation in our application, since it assumes the baseline classifier to be of good quality, while by default it should be considered to be severely suboptimal.

4.3.3 Iterative DARCA

Different from the previous strategies, here, we wish to mimic the active learning domain adaptation [120], where at each iteration, RCA chooses n-subjects from the target domain to fine-tune the baseline networks (see Figure 4.2-3). At the first iteration, we fine-tune the baseline network with best-5 subjects selected by RCA. This new network is used to segment the test images from target domain for which accuracy is again predicted using RCA. At the second iteration, we fine-tune the network again with the new best-5 and worst-5 selected subjects.

The results in Table 4.2 show that the second iteration with worst-5 subjects gives higher accuracies than fine-tuning with best-5 RCA. The combination of best-5 (1st iteration), and worst-5 (2nd iteration) by RCA performs almost the same (DSC: 0.828) as fine-tuning using all of the labeled target data (DSC: 0.830). Additionally, the 2nd iteration with worst-5 RCA selection generally improves the 1st iteration (by best-5 RCA selection) accuracies. Hence, with less labeled data we can save time with similar results.

4.3.4 DARCA on Right Kidney Segmentation

From the three different strategies of DARCA in liver segmentation, we can see that fine-tuning with DARCA gives better results, less time-consuming (compared to training from scratch),

Table 4.2: Strategies of DARCA on liver segmentation: Training from scratch, fine-tuning, and iterative scheme (only for 2nd iteration with RCA selection) with different choice of subject selection.

Strategies of sample selection	**Training from scratch**	**Fine tuning**	**Iterative**
Baseline	0.639 (0.149)	0.639 (0.149)	0.639 (0.149)
All T	0.831 (0.074)	0.830 (0.066)	n/a
Random 5	0.720 (0.103)	0.710 (0.172)	n/a
Worst 5 (real)	0.797 (0.051)	0.619 (0.256)	n/a
Worst 5 (RCA)	0.799 (0.048)	0.771 (0.156)	**0.828** (0.072)
Best 5 (real)	0.747 (0.152)	0.723 (0.173)	n/a
Best 5 (RCA)	0.755 (0.148)	0.687 (0.191)	0.777 (0.107)
Random 10	n/a	0.765 (0.146)	n/a
Worst 10 (real)	n/a	0.745 (0.186)	n/a
Best 10 (real)	n/a	0.779 (0.175)	n/a
Best 5 and Worst 5 (real)	0.823 (0.058)	**0.842** (0.050)	n/a
Best 5 and Worst 5 (RCA)	**0.831** (0.064)	0.835 (0.065)	n/a

Table 4.3: Paired t-tests are performed for significance analysis against **Best-5 Worst-5** sample selection strategy by RCA, with the significance threshold at p= 0.05

Sample selection	**Mean DSC**	**Variance**	**p-value**
Baseline	0.6390	0.0222	1.0383E-10
Random 10	0.7654	0.0213	0.0018
Worst 10	0.7446	0.0347	0.0008
Best 10	0.7799	0.0309	0.0326
Best-5 Worst-5	0.8346	0.0042	*

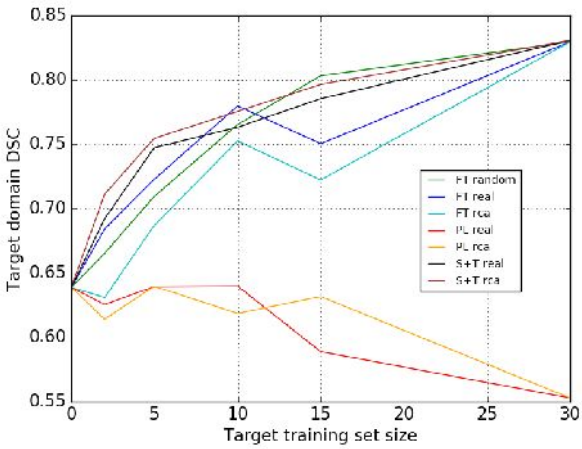


Figure 4.3: Plot for different n-selection training size in different strategies: Fine-tuning (FT), fine-tuning with pseudo-labels (PL), and training from scratch (S+T). Similar trends are shown between real and RCA selection on FT and S+T with different size.

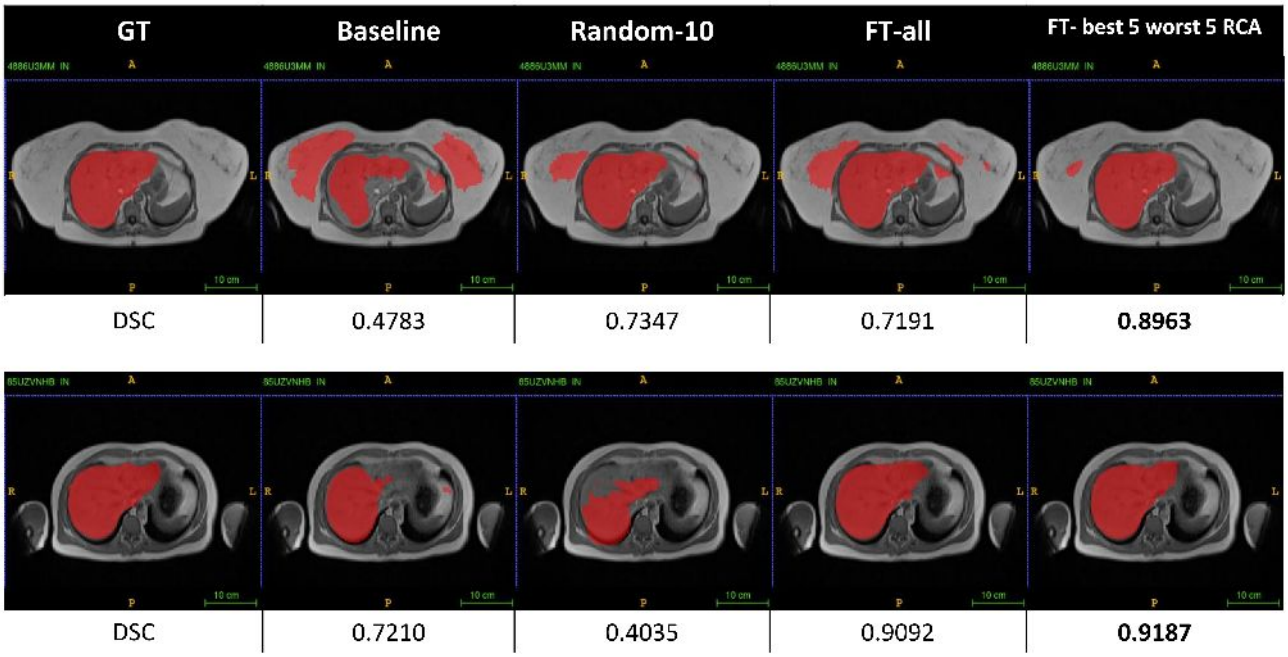


Figure 4.4: DARCA fine-tuning in liver segmentation. Combination of lowest and highest RCA prediction can give a better result than fine-tuning with random selection and with all of the target subjects.

and no iterative scheme needed. Also, from the results in liver segmentation (Table 4.2), combination of best-5 and worst-5 subjects always gives better or similar results than using all of the subjects from domain T, in all strategies. To validate these results, we also explore DARCA-FT in a different task: right kidney segmentation.

When the network is trained and tested for right kidney segmentation on one domain (target dataset), it gives the highest performance. We set this as the upper-bound, as shown in Table 4.4. The performance drops significantly when using a trained network on source dataset and test it on target dataset, we set this as the baseline (see Table 4.4).

The result (DSC: 0.716 with RCA selection) is better than fine-tuning with all of the subjects from T (DSC: 0.658), and similar to when we train from scratch using all target subjects (see Table 4.5). This means we could cut the processing and annotation time. Figure 4.5 depicts some examples on how DARCA-FT with combination of best-5 and worst-5 subject selection improves the baseline and gives better segmentation accuracies than fine-tuning with full size labeled target datasets.

Table 4.4: Baseline and upper-bound accuracies for right kidney segmentation

Training	**DSC (mean (stdv))**
Train on S (baseline)	0.417 (0.263)
Train on T (upper-bound)	0.813 (0.098)

Table 4.5: DARCA-FT on right kidney segmentation. Combination of best and worst subject selection gives the best result, and RCA selection also gives a similar accuracy to the real selection.

Strategies	**Fine tuning (mean (stdv))**
Baseline	0.417 (0.263)
Training from scratch (all T)	0.719 (0.106)
FT all T	0.658 (0.114)
FT Random 5	0.506 (0.278)
FT worst 5 (real)	0.416 (0.254)
FT worst 5 (RCA)	0.358 (0.274)
FT best 5 (real)	0.500 (0.293)
FT best 5(RCA)	0.421 (0.319)
Best 5 and worst 5 (Real)	**0.726** (0.126)
Best 5 and worst 5 (RCA)	0.716 (0.122)

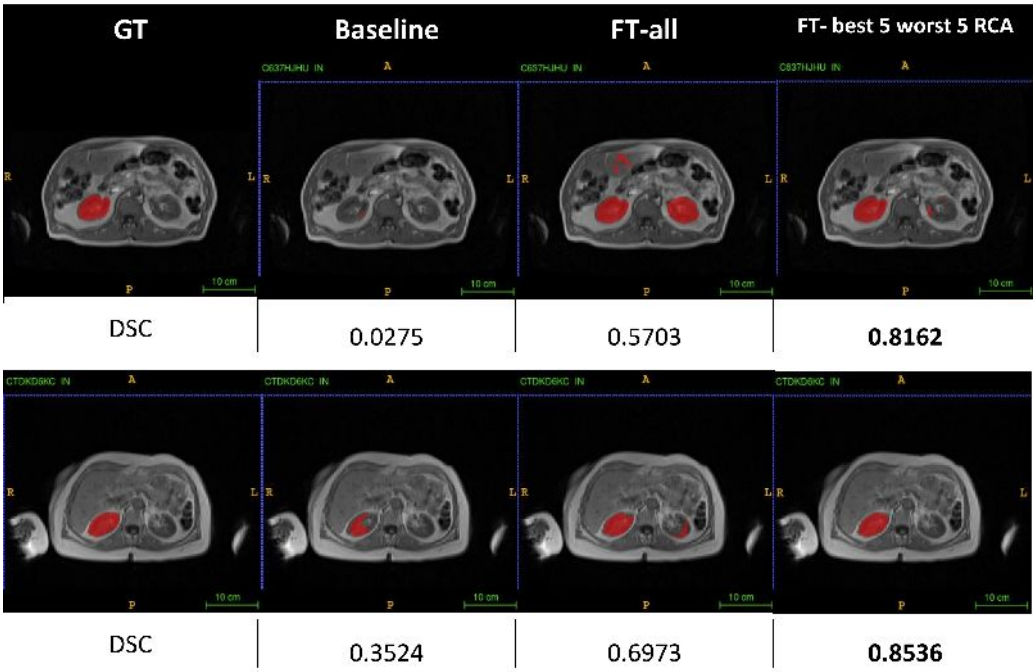


Figure 4.5: DARCA fine-tuning in right kidney segmentation. Combination of lowest and highest RCA prediction can give a better result than fine-tuning with all of the target subjects.

4.4 Discussion and Conclusion

Set size and subject selection are important in domain adaptation, where usually labels are not available in one of the domains. Thus, we have explored whether it is useful to select only the "valuable" subjects by RCA to be manually annotated.

Pseudo-labels, which normally are used in semi-supervised learning and regularization, seem not very useful in *supervised* DARCA. As observed in Figure 4.3 the performance drops as we increase the number of pseudo-labels in fine-tuning. Pseudo-labels will introduce more noise confusing the training of the networks and make them incapable to be applied to the new domain.

We have demonstrated that DARCA with few labeled data can perform similarly and/or better to full-size labeled target data. In the examples of Figure 4.4 and Figure 4.5, Table 4.2 and Table 4.5, DARCA with best-5 and worst-5 subjects show consistent results across different tasks (liver and kidney segmentation).

In the case of real DSCs between [0.6, 0.8), RCA underestimates the predictions [140]. This led to a different subject selection. In the future, an improvement of RCA prediction for medium level DSCs [0.6-0.8) needs to be investigated so that it can work more accurately.

Our study only focuses at a predefined number of selected subjects, and a more thorough exploration needs to be done in future work. Traditional active learning may have more flexibility regarding the number of "valuable" samples to be included, but it requires iterations, which is time consuming. DARCA only needs RCA to predict the DSCs from the baseline segmentation where we can select the highest and lowest predicted subjects to be incorporated in a quick fine-tuning procedure.

Therefore, we can conclude that DARCA could save processing time (no iterations needed) and annotation time with promising results avoiding the need for a large annotated target database.

Chapter 5

Learning between Imaging Modalities

This chapter is based on:

Vanya V. Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker. Multi-Modal Learning from Unpaired Images: Application to Multi-Organ Segmentation in CT and MRI. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 547-556, 2018.

In whole-body imaging, not only we need to learn between different scanners (as explained in the previous chapter), but also we need to learn between different imaging modalities. It is because in clinical practice, different imaging modalities (MRI, CT, Ultrasound, etc.) may be used to acquired the same structures.

Convolutional neural networks have been widely used in medical image segmentation. The amount of training data strongly determines the overall performance. Most approaches are applied for a single imaging modality, e.g., brain MRI. In practice, it is often difficult to acquire sufficient training data of a certain imaging modality. The same anatomical structures, however, may be visible in different modalities such as major organs on abdominal CT and MRI.

In this chapter, we investigate the effectiveness of learning from multiple modalities to improve the segmentation accuracy on each individual modality. We study the feasibility of using a dual-

stream encoder-decoder architecture to learn modality-independent, and thus, generalisable and robust features. All of our MRI and CT data are *unpaired*, which means they are obtained from different subjects and not registered to each other. Experiments show that multi-modal learning can improve overall accuracy over modality-specific training. Results demonstrate that information across modalities can in particular improve performance on varying structures such as the spleen.

5.1 Introduction

In clinical practice, multiple imaging modalities are used to capture anatomical structures such as major abdominal organs. For example, CT and MRI both give clear images of organs such as the liver, the heart, or the kidneys, however, the visual appearance is completely different due to the very different underlying physical principles of each imaging technique.

For the human eye, however, it is quite easy to identify the same structures in different modalities. For example, after showing an abdominal CT scan and pointing out where the liver is, even a non-expert will be able to find the liver in an MRI scan. For algorithms trained to segment livers in CT, however, it is quasi impossible to segment the livers in MRI due to the different nature of the imaging features that are learned in a modality-specific setting. In this chapter, we explore the power of multi-modal learning to extract abstract representations that are modality-independent and have the potential to increase accuracy and robustness when employed as features in a supervised setting.

Deep-learning based approaches such as convolutional neural networks (CNNs) have become popular in medical imaging but they require large amounts of data to achieve good results, near human-level capability [8, 105]. However, in practice, it is often difficult to acquire sufficiently many manually annotated datasets from one modality. Hence, there is a strong desire to utilise all available data for the same or similar tasks even if the images are from different modalities. The main question is whether the information from one modality can actually improve the performance of a task on another modality [9]. We aim to answer this question through an

extensive set of experiments with novel CNN architectures for multi-modal learning.

Most works in medical image segmentation refer to the term *multi-modality* when using different sequences in MRI, such as T1-weighted, T2-weighted, diffusion, or functional MRI. The contrast and information in each sequence is different. There are varying approaches to these multi-modal segmentation problems, even though the most common approach is by concatenating the inputs from different modalities as different input channels to the networks [61, 74, 84]. Another option is to train one CNN for each modality and fusing multi-modality features from high-layers of each network [103]. High-level representations from different modalities are complementary, and might yield improvement in performance.

In multi-modal brain MRI segmentation, cross-modality convolution with a deep encoder-decoder architecture has been explored by [135] with convolutional LSTM to model the sequential correlations between slices. They show that cross-modality convolution can effectively aggregate the information between modalities to produce better results [135]. These works assume that all modalities are always available for each individual subject.

The Hetero-modal network architecture (HeMIS) [62] and scalable multi-modal CNN (ScaleNet) [42] are built for multi-modal brain MRI segmentation that can deal with missing modalities, however, they assume modalities are from the same subjects and co-registered. In our case, all of the data are *unpaired*, the acquisitions are from different subjects and images are not registered.

All previous works on multi-modal segmentation use *paired* data, meaning that the images are acquired from the same subject and co-registered across modalities. Hence, to the best of our knowledge, learning across unpaired data from different imaging modalities has not been explored yet in medical imaging.

Transfer learning is one way to share knowledge between different domains. Fine-tuning a pre-trained model from natural images is shown to be useful for medical image analysis, even though the source and target domains are very different [133]. Transfer learning between CT and MRI has been explored in [159]. They showed that the accuracy of MRI segmentation

can be improved by using shape priors from CT. In deep learning based segmentation, learning different modalities needs different parameters in order to work well on a particular network architecture [97]. In general, it appears that training each modality separately has so far led to best results, which might be due to the use of non-optimal architectures.

In medical imaging, training multi-task on one network had been conducted by [98]. They showed that the performance of learning multi-tasks and multi-modalities on a single CNN is *equivalent* to a single network trained specifically for one-task and one-modality. Instead of using a single network for multi-task learning, [80] applied a multi-modal encoder-decoder network with shared latent space and shared skip connections. Their results show a potential of shared representations from different modalities to improve the multi-task performance.

Multi-modal learning can also be applied to image synthesis, where a joint representation in a common latent space preserves the local consistency of the images [67]. Using this representation, cross-modality synthesis in MRI (T2 to T1, FLAIR to T2 and vice versa) can be achieved in high-resolution. Multi-modal image synthesis of MRI is explored in [71, 24] by composing the networks of encoders (for each modality), fusing latent representations, and a decoder to produce an output image.

Our work exploits the shared representation in multi-modal learning, particularly in unpaired CT and MR images, which has not been investigated before. We set up different network architectures and multi-modal training schemes to evaluate the effect on segmentation accuracy for each modality.

The contributions of this chapter are as follows:

- We learn modality-independent features from CT and MRI using an encoder-decoder network architecture
- We explore dual-stream networks with different strategies for sharing information between tasks
- We show improved segmentation accuracy of variousorgans when using multi-modal learning

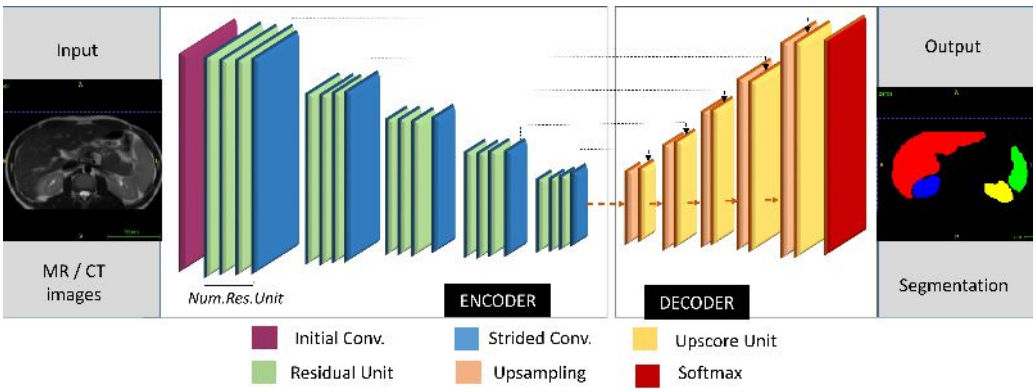


Figure 5.1: Baseline network architecture for multi-organ segmentation: FCN based network with encoder-decoder structure

5.2 Methodology

In this section, we will explain our proposed networks and multi-modal learning framework for multi-organ segmentation on different modalities, MR and CT. Several cases and setting will be detailed, including different network architectures, multi-modal learning on singe-task and multi-task.

Previous multi-modal segmentation approaches [62, 42] learn to fuse the information from modalities and to handle missing modalities. Thus, pairs of n-modalities have to be available at least once in training time. This approach is not feasible when there are only unpaired images from different modalities. Our challenge is to build a model which can take *unpaired* inputs from n-image modalities to produce an accurate segmentation for each modality.

5.2.1 Network architectures

As a baseline, we use a Fully convolutional network (FCN) [125] with residual layers [64] for multi-organ image segmentation (see Figure 5.1). The network can be seen as an encoder-decoder architecture.

Encoder. The network consists of a number of residual feature encoding blocks with pre-activation [64], as shown in Figure 5.2. We specify the number of residual units in each scale, which is chosen experimentally. The residual unit block consist of batch normalisation *(BN)*,

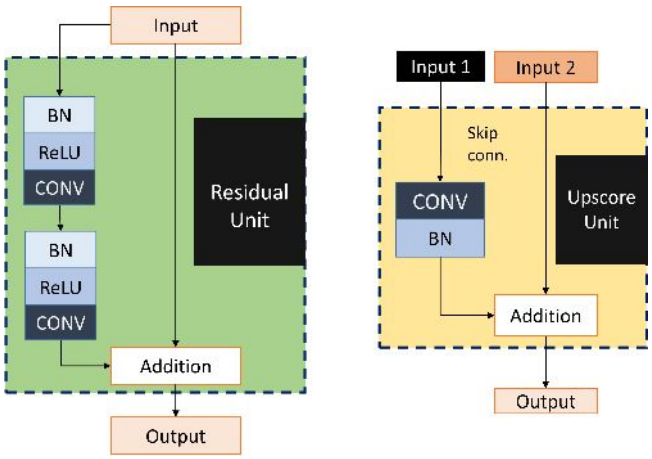


Figure 5.2: Left: Residual unit with pre-activation. Right: Upscore Unit. Input 1: Higher resolution features from the encoder (same scale) to add. Input 2: Input features from decoder to be upscored.

activation function *(ReLU)*, followed by 3D convolutional layers *(Conv)* for extracting image features. To handle the stride convolution, we add pooling to the input before the addition in residual unit. Then, we use padding if the number of filters are different between input and output. Before moving on to the next scale, the features are downsampled via strided convolutions. We use kernel size 3^3 with stride 1 and padding size 1 for all convolutional layer, and kernel size twice the stride for strided convolution.

Shared representation. Multi-modal learning focuses on gaining shared representations from multi-modal data. Since our data is unpaired, straightforward concatenation of extracted features from different modalities is not reasonable. Moreover, at test time we want to be able to perform segmentation on a single modality acquired for a single subject. We aim to improve the shared representation by using an encoder-decoder architecture, to investigate which part of encoder or decoder should be shared across modalities.

Decoder. The decoding stage uses fully convolutional layers to map a target output modality. Feature maps learnt at different scales are upsampled to the original resolution. The decoder consists of upscore units as depicted in Figure 5.2 with kernel 1^3. The input is upsampled before going into the upscore unit. Inside the upscore unit, the features from the encoder are learnt to produce a sparse feature map. We apply skip connections, with the 3D convolution of the encoder features followed by batch normalisation [64]. This is added to the input features

to be upscored.

Finally, the rescaled output is fed into a softmax layer to produce the probabilistic label map. The highest softmax probability for each class is computed to yield the final segmentation result.

5.2.2 Dual-stream

To effectively infer the multi-modal features from CT and MRI, we propose a dual-stream network architecture [1]. By using individual streams for each modality, we can investigate when it is best to merge or split the streams. Dual-stream also allows to process *unpaired* images in one network, which is not possible with standard multi-channel architectures [74, 62] for *paired* data. The images from unpaired multi-modal data have no correspondences which makes it difficult to find correlations across modalities. To exploit the commonality among different modalities, streams of encoder/decoder are connected with each other via the shared latent representation.

5.2.3 Dual-stream Networks for Multi-modal learning on unpaired images

For the dual-stream model, we also implement an architecture with skip connections to propagate features from encoders to decoders. In the basic network in Figure 5.1, the skip connections are used to recover the full spatial resolution at the network output, by merging features skipped from the various resolution levels on the encoders. We still retain the skip connections in all of the versions of dual-stream networks, because we still need the features from the encoders regardless of the different stream as a part of multi-modal learning.

Skip connections have been shown to be effective to learn the shared latent representation from different encoders in multi-modal architecture [80]. If given one input modality, the encoder

[1] Source code available at `https://github.com/vanya2v/Multi-modal-learning`

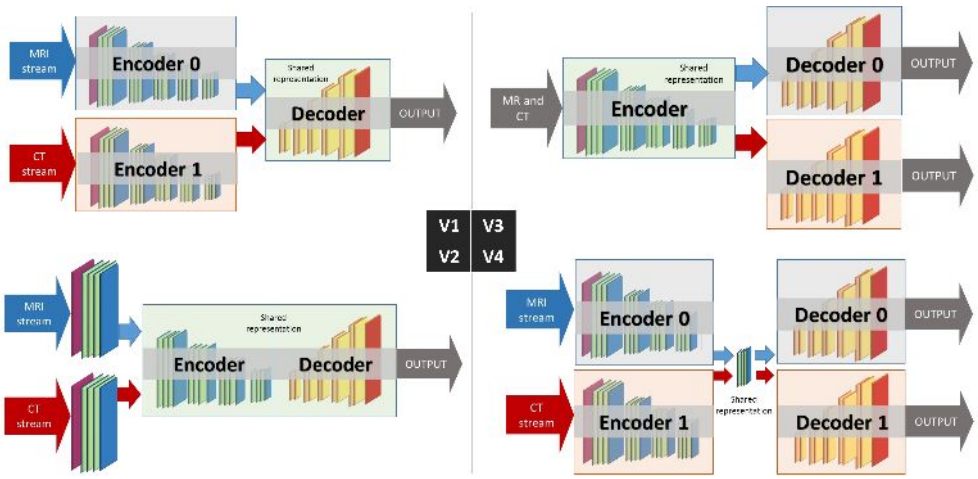


Figure 5.3: Dual-stream models. Version 1 - modality-specific encoder and shared decoder. Version 2 - shared part-of-encoder and decoder. Version 3 - shared encoder and modality-specific decoder. Version 4 - different streams in encoder, shared last layers of encoder, and modality-specific decoder.

generates a representation, which is then shared via skip connection and decoded through different decoders. Hence, we also use features that are not yet shared, which can be helpful to support the segmentation from multi-modalities.

The FCN-based architecture is the most suitable choice for dual-streams as the features can be merged from different scales in the encoder which vary in semantic information. Implementing the dual-stream model is slightly more complex in U-Net [115] and DeepMedic architectures [74]. For U-Net, the encoder feature maps needs to be *concatenated* with the upsampled feature maps from the decoder at every scale. For DeepMedic, the multi-scale nature of the network needs to be taken into account. Hence, for this work we opt to use an FCN-based model in all our experiments.

5.3 Experiments and Results

In this section, we evaluate the multi-modal encoder-decoder networks for multi-organ segmentation using two datasets: MRI and CT data. The implementation details are below.

5.3.1 FCN-based dual-stream Networks

Data and pre-processing

We use two datasets of 3D abdominal images from MRI and CT [82] to evaluate the models. We train the model with a roughly balanced dataset: 34 subjects for MRI and 30 subjects for CT to keep it agnostic to the modality intensity distribution. For both CT and MRI data we split the data into 2-folds for cross-validation. From the MRI database we used T2-weighted images. Both CT and MRI scans have the same four organs manually annotated: liver, spleen, right kidney, and left kidney delineated by clinical experts. All of the CT and MRI data are unpaired, which means that the acquisitions are from different subjects. All experiments are carried out using NVIDIA GPUs.

We perform all experiments on resampled data with 2 mm voxel spacing. The original abdominal scans of MRI and CT capture the area from neck to knee, with varying image sizes according to patient body size. In order to have similar FOV (Field of View) between all of the images, we crop the volumes so that they only cover all four organs to be segmented - removing uninformative areas. We take specific preprocessing measures for CT images to account for the differences in intensity distribution. We clip all values lower than specified threshold, before zero-meanunit-variance intensity normalisation. We add Gaussian noise and intensity offsets to the images for data augmentation, as a common approach in training the deep networks. Then, we use weighted class sampling during training to compensate for class-imbalance.

Training Details

We implement the model using DLTK (Deep Learning Toolkit) [110], with a Tensorflow backend [1]. We use the DLTK FCN segmentation model as a baseline network, which has been optimised for abdominal multi-organ segmentation. We experiment with both cross-entropy and Dice-loss, and choose the Dice-loss for the better training stability. Let $o_i \epsilon [0, 1]$ be the i the output of the network and let $y_i \epsilon [0, 1]$ be the corresponding label. The Dice-loss is then defined as follows:

$$L_{Dice} = \frac{2\sum_i o_i y_i}{\sum_i o_i + \sum_i y_i}$$

The model is trained using Adam optimisation with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 10^{-5}$. We use mini-batch training of 16 training examples with size 64^3, to provide enough context at 2 mm resolution. The number of filters for each scale is 16, 64, 128, 256, and 512. Training iteration is for 10k steps and the training examples are shuffled from all training images. However, for multi-modal learning, training examples are drawn in an alternating way: one iteration contains only MRI data, and the subsequent iteration contains only CT data. Since we use alternating batches from modalities in training, during testing procedure we can simply feed single modality data for inference.

Evaluation of segmentation performance

We repeat the training and testing procedures for all of the models three times, to average out random variations. The results presented are the total of all experiments and the average accuracies from all scores. Results are evaluated using the Dice score for the four individual abdominal organs: liver, spleen, right kidney, and left kidney.

To demonstrate the benefit of multi-modal learning, we extensively compare several different approaches and baselines as detailed below.

Baseline multi-organ segmentation model

The segmentation model in Figure 5.1 is used for modality-specific segmentation. Here, we train the network on MRI data and only tested on MRI data, and the same procedure is done for CT data. The number of residual units is 3 per scale and the learning rate is 0.001, which are chosen via experimentation. As a baseline, one network is trained separately for each imaging modality.

Joint learning

The same model as in Figure 5.1 with the same parameters, network capacity, and optimisation is trained for a combination of CT and MR data. The only difference in this joint learning is the data, we train a single network with input from *both* CT and MRI data to perform the segmentation task. Additionally, we use alternating batches in training phase, one iteration for MRI batch and the next iteration fetched the CT batch. Hence, the network can learn from balanced examples from both modalities.

Fine-tuning (FT)

A straight forward approach to solve the cross modal segmentation is by a traditional fine-tuning [124, 133]. Here, a pre-trained network on one modality is fine-tuned to fit another modality, as we can see different modalities as different domains. For multi-modal fine-tuning, we fine-tune the layers of pre-trained network from MR for CT segmentation, and vice versa. Thus, the weights is adjusted for multi-modal segmentation.

Dual-stream

Dual-stream models are built to handle multi-modal segmentation in one network, with one stream per modality. If the input is from MRI data then stream 0 will be chosen, while stream 1 will be activated if the network has CT input data. In the training phase, we also apply alternating batches as in joint learning, and the loss function is back-propagated per modality. The optimisation in all dual-stream networks are the same to provide the lowest error and a fair comparison. We still use the same hyper-parameter settings as in the baseline model, but with a lower learning rate 0.0001 because of a more complex architecture. Different dual-stream versions implemented in our experiment are illustrated in Figure 5.3.

Version 1 (V1). This model has a separate encoder for each modality. MRI data will go to encoder 0 and CT data will go into encoder 1, and share the same decoder. As in [62] and [42], this type of independent, modality-specific in the back-end and fusing in the front-end model

is shown to work with paired multi-modal segmentation. The decoding task in this model is expected to benefit from high-resolution representations, as the decoder is connected from different encoders.

Version 2 (V2). Both modalities share most part of encoder, and decoder. A separate stream for CT and MRI is only applied in the input and residual units at the first scale. Then, they share the same encoder and decoder part. We would see that if low level layer separation is not enough for modality-specific feature representation.

Version 3 (V3). CT and MRI share the same encoder but have different streams in decoding part. By having both modalities share the same encoder, they expect to generate a single shared representation to discover correlation across the modalities, which is then modality-specific decoded. This model is similar to [101] when the model only has one modality to reconstruct both modalities. In our case, the segmentation output is for each modality. Separate decoders in the network is shown to be helpful because it will try to reconstruct to specific modality and force to recover the input for each modality.

Version 4 (V4). Inspired by one of the architectures in bi-modal deep AE [101] and multi-modal encoder-decoder networks in [80], we use dual-stream in both encoder and decoder. Both MR and CT flows into a different stream in the encoding part, but only share weights at the last scale of encoder. The rationale behind this is that high resolution layers have high-level-like shape/structures that may have more benefit for same-task multi-modal learning. These layers can be seen as a shared latent representation which is then fed into a modality-specific decoder. In other words, all encoder/decoder streams are connected with the shared latent representation.

Results

Multi-modal learning can leverage the shared information in both modalities in one pass, unlike the traditional transfer learning with the sequential training (initial training and fine-tuning the pretrained networks). Additionally, from our experiments, we found that fine-tuning did not improve over models trained jointly and individually.

We compare the performance of multi-modal learning (joint and dual-stream versions) to individual learning in MR and CT data. Because we ran the benchmark procedures (training and testing in 2-fold cross-validation for each modality) in three repetitions for each model, we present the results in Table 5.1, 5.2, 5.3 as the average scores from all experiments.

We report how the multi-modal learning affects the Dice scores of each organ segmentation for different modalities. For MRI, in Table 5.1, we can observe that multi-modal learning by dual-stream V4 generally improves the performance for all organs, compared to individual segmentation. From Table 5.1, a better accuracy is achieved in spleen and left kidney segmentation. Note that there is a significant improvement in right kidney segmentation, by four Dice scores. Joint CT-MR also gives a slight increase in the right kidney accuracy.

Table 5.2 shows the benefit of multi-modal learning in CT data with dual-stream V4. It is interesting to note that the multi-modal learning is more helpful to segment organs in CT (than in MR), as shown for left kidney where all multi-modal models (joint and dual-stream models) improve the individual performance. A high rise of Dice scores is shown in small organs, especially spleen and left kidney. Joint CT-MR also gives slightly better accuracy in spleen segmentation. In both modalities, the accuracy of liver segmentation with multi-modal learning is consistent, as individually trained segmentation already gives a high accuracy.

In terms of overall performance, we report the results for all organ segmentation of benchmark models in Table 5.3. Dual-stream V4 outperforms the individual training with almost three Dice scores increment. However, other multi-modal learning models fail to improve the segmentation of individual learning. Despite of its lower accuracy than dual-stream V4, joint CT-MR can still train a single neural network to do quite well on both modalities. It is because the network effectively compresses the information of both databases into a model that usually fits one modality and improves the performance. We did further experiments to make the network's capacity larger, but we did not obtain better results in joint CT-MR. Fine-tuning the networks to work on another modality is shown to be less effective and less accurate than joint learning and multi-modal learning with dual-stream V4. Moreover, fine-tuning require a separate network to be trained, which is more time consuming.

Table 5.1: Multi-modal learning in MR images

Organ	Individual	FT	Joint	V1	V2	V3	V4
Liver	0.913	0.876	0.903	0.877	0.862	0.905	**0.914**
Spleen	0.772	0.788	0.772	0.728	0.732	0.736	**0.790**
Right Kidney	0.835	0.788	0.842	0.804	0.792	0.819	**0.871**
Left Kidney	0.821	0.819	0.771	0.748	0.741	0.810	**0.833**

Table 5.2: Multi-modal learning in CT images

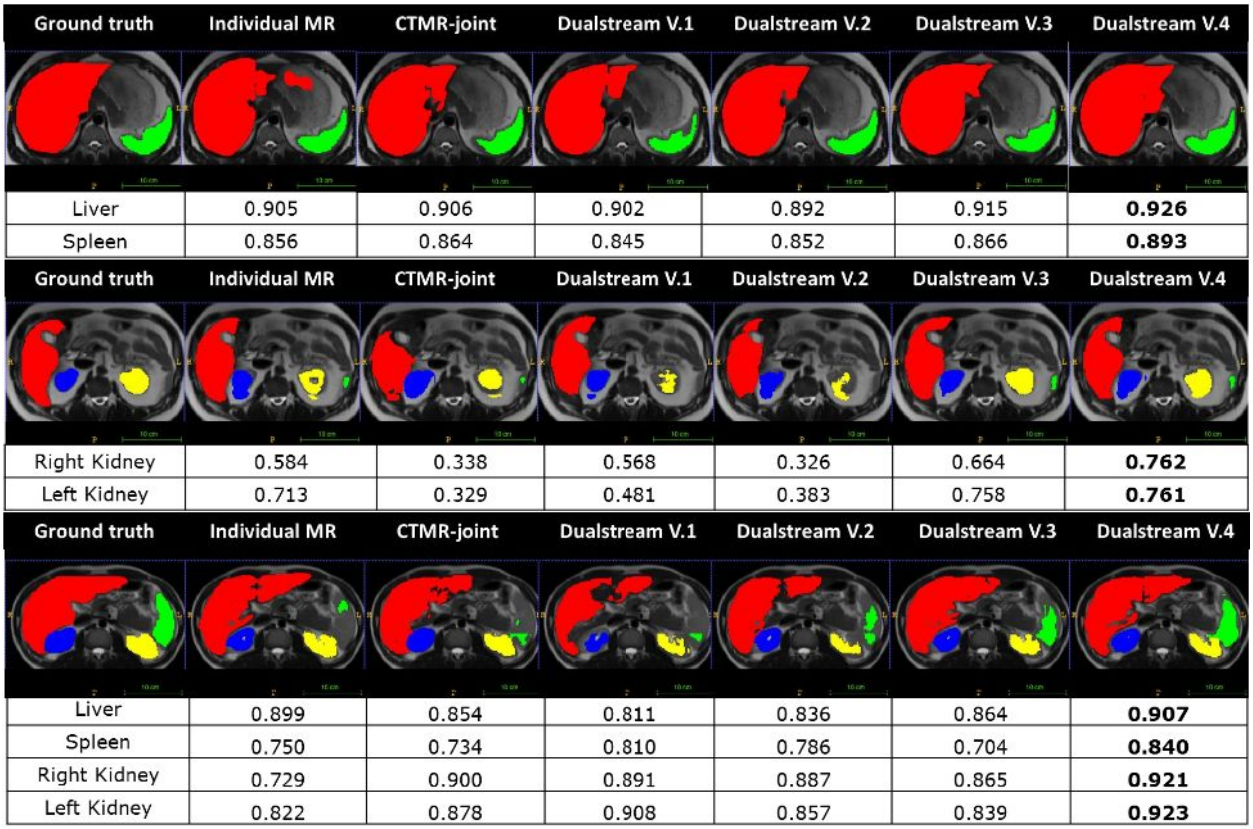
Organ	Individual	FT	Joint	V1	V2	V3	V4
Liver	0.914	0.836	0.875	0.900	0.902	0.895	**0.919**
Spleen	0.824	0.729	0.834	0.822	0.824	0.817	**0.859**
Right Kidney	0.820	0.682	0.806	0.824	0.815	0.793	**0.838**
Left Kidney	0.808	0.679	0.822	0.833	0.818	0.819	**0.851**

From Table 5.3, a shared decoder (as in dual-stream V1 and V2) cannot effectively leverage the multi-modal features for segmentation. In dual-stream V1, we tried to combine the high-level features from both modalities in a shared decoding phase. In dual-stream V2, we only extract low-level features in different streams, to be encoded and decoded together. In contrast to shared-decoder in *paired* multi-modal data [42, 62], this model cannot learn effectively in *unpaired* data. Dual-stream V1 and V2 are forced to have the same decoding phase. Meanwhile, dual-stream V3 and V4 give better performance than V1 and V2 because of their modality-specific decoders. It is interesting to note that different streams in upscoring could increase the multi-modal learning capability, as shown in [101, 80]. However, in dual-stream V3, the encoders are tied between modalities, making it harder to leverage the multi-modal features. As a result, dual-stream V4 with split streams of encoder, shared latent representation (last layers on encoder), and modality-specific decoder gives overall best performance.

By visual inspection, we can see how multi-modal learning can improve multi-organ segmentation. For example in MRI data in Figure 5.4 (top), multi-modal learning can reduce the false positive of liver segmentation and increase spleen accuracy. In Figure 5.4 (middle), both kidneys give better accuracies in dual-stream V4. Overall organ segmentation in Figure 5.4 (bottom) is improved by dual-stream V4.

Table 5.3: Overview on multi-modal learning on all organs

All organs	Individual	FT	Joint	V1	V2	V3	V4
Average	0.838	0.774	0.828	0.817	0.811	0.824	**0.860**
Std.Deviation	0.019	0.058	0.021	0.059	0.065	0.034	**0.011**



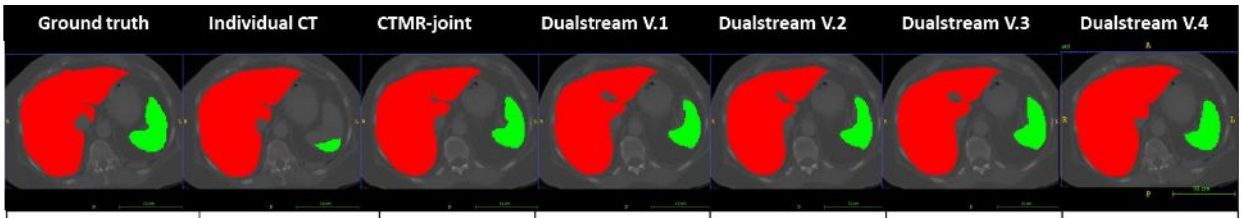
	Individual MR	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
Liver	0.905	0.906	0.902	0.892	0.915	**0.926**
Spleen	0.856	0.864	0.845	0.852	0.866	**0.893**
Right Kidney	0.584	0.338	0.568	0.326	0.664	**0.762**
Left Kidney	0.713	0.329	0.481	0.383	0.758	**0.761**
Liver	0.899	0.854	0.811	0.836	0.864	**0.907**
Spleen	0.750	0.734	0.810	0.786	0.704	**0.840**
Right Kidney	0.729	0.900	0.891	0.887	0.865	**0.921**
Left Kidney	0.822	0.878	0.908	0.857	0.839	**0.923**

Figure 5.4: Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow).

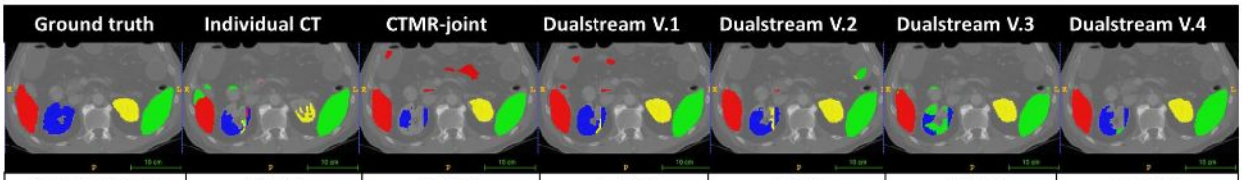
We also visualise how multi-modal learning can help the segmentation in CT. In Figure 5.5 (top), we can see how all multi-modal learning (CT-MR joint, dual-stream V1, V2, V3 and V4) can help spleen segmentation. Dual-stream V4 segmentation of left kidney has a significant increase in Dice scores (Figure 5.5-middle), compared to individually trained segmentation. Multi-organ segmentation is improved by dual-stream V4 in Figure 5.5 (bottom).

Box-plots in Figure 5.6 and 5.7 from all experiments illustrate the benefit of multi-modal learning. Dual-stream V4 gives an improved segmentation performance compared to individual learning, particularly in kidneys and spleen. In dual-stream V4, the outliers are reduced as well as the variance on Dice scores, for both modalities.

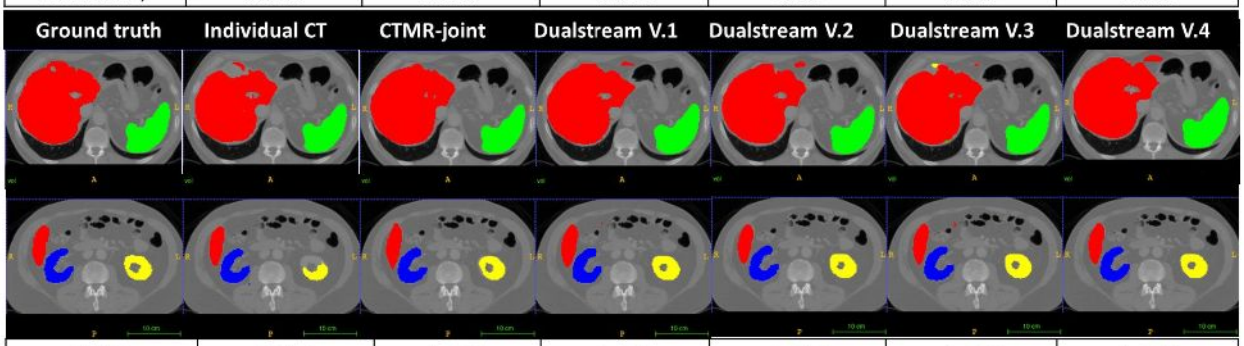
We then evaluate the performance of dual-stream V4 if the stream is swapped during testing, but only in one experiment (not in three repetitions for each model, as in the results in Table 5.1, 5.2, 5.3 with the average scores from all experiments). Table 5.4 gives the summary of



Spleen	0.384	0.891	0.846	0.836	0.904	**0.907**



Right Kidney	0.818	0.783	0.869	0.842	0.735	**0.898**
Left Kidney	0.723	0.867	0.912	0.855	0.893	**0.937**



Liver	0.931	0.918	0.931	0.934	0.936	**0.943**
Spleen	0.942	0.908	0.940	0.934	0.930	**0.951**
Right Kidney	0.935	0.921	0.928	0.936	0.932	**0.939**
Left Kidney	0.743	0.902	0.919	0.916	0.918	**0.935**

Figure 5.5: Multi-modal learning on CT. Liver (red), spleen (green), right kidney (blue), left kidney (yellow).

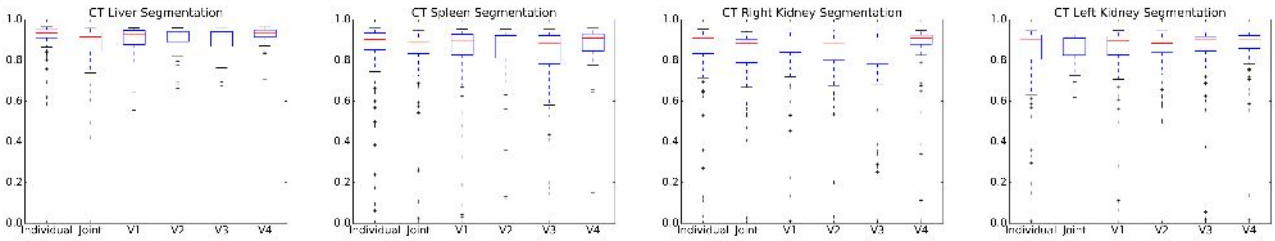


Figure 5.6: Boxplot of different organ segmentations on CT data: liver, spleen, right kidney, and left kidney segmentation (left-right). Different models on x-axis and Dice scores on y-axis.

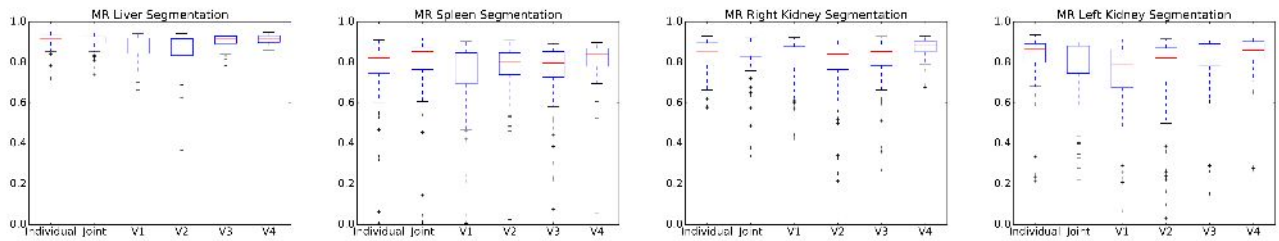


Figure 5.7: Boxplot of different organ segmentations on MRI data: liver, spleen, right kidney, and left kidney segmentation (left-right). Different models on x-axis and Dice scores on y-axis.

Table 5.4: Swap stream of dual-stream Version 4

	MR			CT		
Organ	Individual	V4	V4-Swap	Individual	V4	V4-Swap
Liver	0.905	0.911	0.777	0.927	0.929	0.859
Spleen	0.783	0.787	0.667	0.855	0.875	0.688
Right Kidney	0.834	0.875	0.728	0.820	0.843	0.611
Left Kidney	0.818	0.821	0.593	0.763	0.866	0.431

the swap-stream in dual-stream V4 for one fold of the experiments. In training, stream 0 is dedicated for MRI data while stream 1 is for CT data. When we swap the stream, the network obviously completely fails, resulting worse results than the individual segmentation. This shows that the feature encoding and decoding parts are dependent on the streams. In contrast, with joint CT-MR, all parts of the network are shared between CT and MR features.

As an additional experiment, we also tried random input on dual-stream networks for one experiment (not in three repetition). The default dual-stream network training is alternate-training: one iteration for MRI and the next iteration is for CT, hence we have relatively balance samples from both modalities. However, when we have random input (not alternating MRI or CT input selection), the results (as shown in Table 5.5) are not very different.

From Table 5.5 we still can see that dual-stream V4 with alternate training gives the best results. From this experiment, we can conclude that if we have imbalance datasets from either modality, dual-stream networks can still work well. Hence, if for example, one hospital has less CT datasets, dual-stream network can still be beneficial, although trained with more MRI dataset, vice versa.

From [98] we learn that multi-task segmentation on a single network performs similarly on

Table 5.5: Random input for training dual-stream network

Organ	Individual	V1	V1-random	V2	V2-random	V3	V3-random	V4	V4-random
Liver	0.916	0.884	0.907	0.854	0.889	0.910	0.894	0.914	**0.916**
Spleen	0.794	0.777	0.802	0.754	0.769	0.791	0.810	**0.818**	0.810
Right Kidney	0.834	0.825	0.821	0.763	0.814	0.828	0.838	**0.855**	0.834
Left Kidney	0.827	0.785	0.814	0.731	0.787	0.835	**0.851**	0.849	0.843
Avg.	0.843	0.818	0.836	0.776	0.815	0.841	0.848	**0.859**	0.851

multi-modal data. From our experiments, we believe that multi-modal learning in a single task (multi-organ segmentation) can benefit from sharing information, even though the data are *unpaired*.

5.3.2 U-Net based Dual-stream Networks

We also conduct the same experiments with dual-stream on U-Net based networks [115]. U-Net is interesting because the architecture is designed for processing biomedical images with an augmentation work-flow. U-Net also applies an architecture with encoder-decoder structures as shown in Figure 5.1, which can be used for supervised tasks like object segmentation.

The difference between FCN and U-Net is on the upscore unit in decoder part. For FCN, as seen in Figure 5.2, input 1 (after convolution and batch-norm) is *added* to input 2. But, for U-Net, instead of adding up, input 1 is simply *concatenated* with input 2 (see Figure 5.8).

Both FCN and U-Net utilise fully convolutional networks, where the latter layers have features that help explaining what the image contains. Both also uses deconvolution networks for upsampling. However, the accuracies for multi-organ segmentation are different between FCN and U-Net based. We compare the performance of multi-modal learning (joint and dual-stream versions) to individual learning in MRI and CT data. Since we run the benchmark procedures (2-fold cross-validation for each modality) in three repetitions for each model, we present the results in Table 5.6, 5.7, 5.8 as the average scores from all experiments.

For MRI, in Table 5.6, we can observe that multi-modal learning by dual-stream V3 improves the segmentation on two organs (liver and left kidney) while on other two organs (spleen and right kidney), multi-modal learning cannot improve the individual segmentation. However,

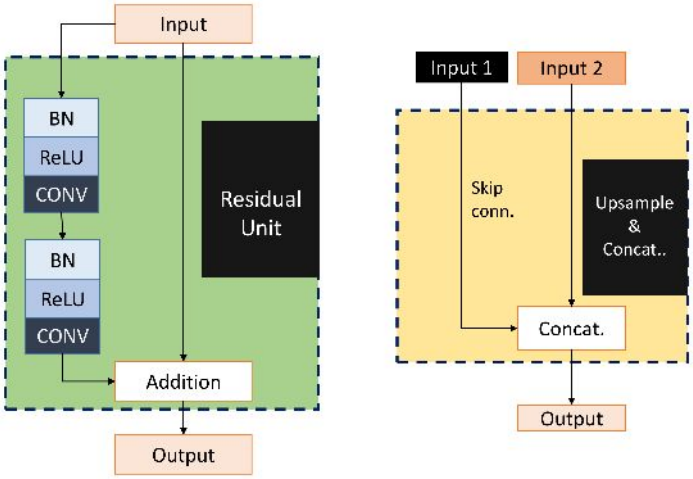


Figure 5.8: Left: Residual unit with pre-activation. Right: Upsample and concat Unit. Input 1: Higher resolution features from the encoder (same scale) to add. Input 2: Input features from decoder to be upsampled and concatenated.

Table 5.6: Multi-modal learning in MR images

Organ	Individual	Joint	V1	V2	V3	V4
Liver	0.899	0.846	0.886	0.844	**0.910**	0.788
Spleen	**0.715**	0.712	0.579	0.584	0.661	0.519
Right Kidney	**0.786**	0.750	0.705	0.709	0.770	0.614
Left Kidney	0.776	0.703	0.660	0.615	**0.789**	0.590

the improvement in MR segmentation is not as significant as those in FCN-based multi-modal learning.

Table 5.7 shows the benefit of multi-modal learning in CT data with dual-stream V4. It is interesting to note that multi-modal learning is more helpful to segment organs in CT than in MRI images. Here, segmentation of all organs achieved better accuracies with dual-stream networks than with individual networks. A high rise of Dice scores is shown in small organs, especially kidneys.

Table 5.7: Multi-modal learning in CT images

Organ	Individual	Joint	V1	V2	V3	V4
Liver	0.879	0.831	0.921	0.922	0.919	**0.926**
Spleen	0.797	0.715	0.865	0.856	0.854	**0.868**
Right Kidney	0.729	0.676	**0.860**	0.813	0.845	0.855
Left Kidney	0.770	0.649	0.846	0.848	**0.865**	0.853

Table 5.8: Overview on multi-modal learning on U-Net based on all organs

All organs	Individual	Joint	V1	V2	V3	V4
Average	0.794	0.735	0.790	0.774	**0.827**	0.752
Std.Deviation	0.054	0.027	0.031	0.076	0.015	0.054

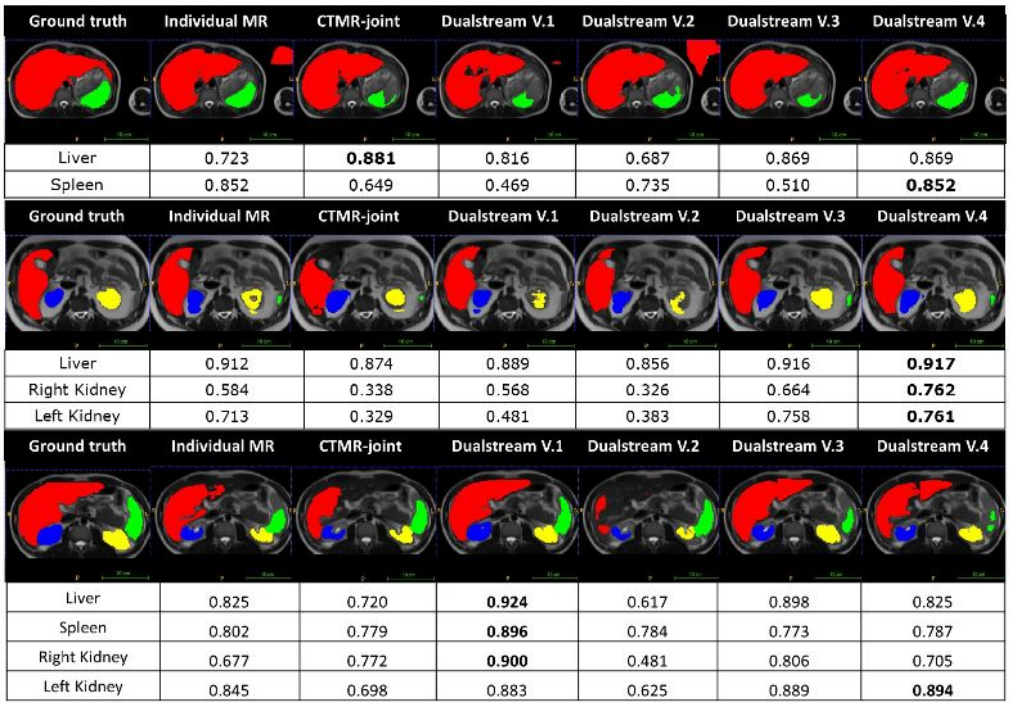


Figure 5.9: Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow).

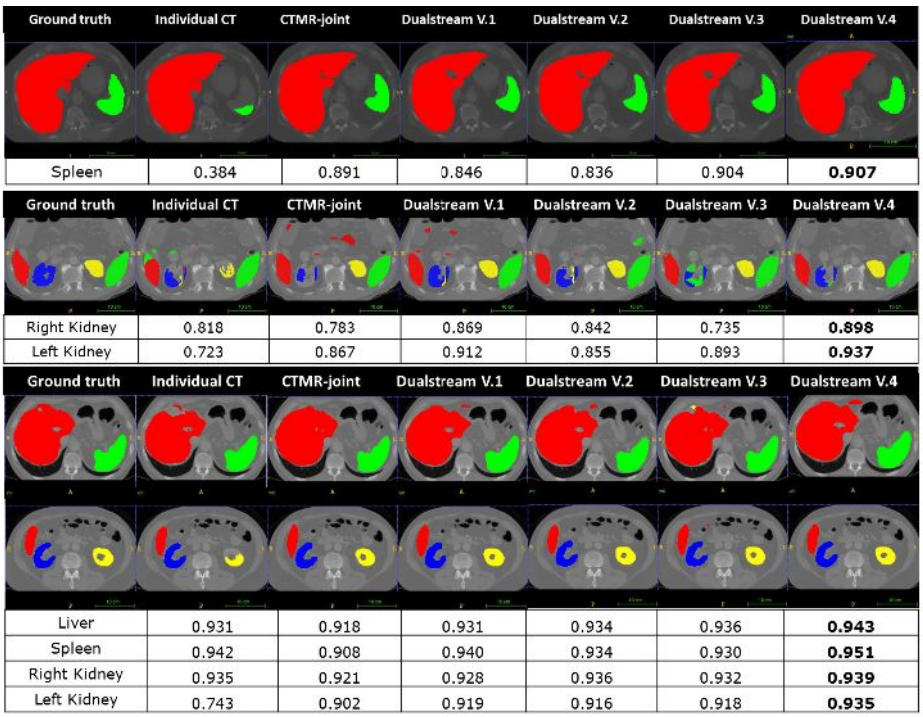


Figure 5.10: Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow).

With U-Net based networks, the best performing strategy depends on the modality. For MRI, dual-stream V3 gives the best results, while dual-stream V4 works best in CT segmentation. However, for the overall results (see Table 5.8) on both modalities, multi-modal learning with dual-stream V3 is shown to give an improvement compare to individual learning, with a 3% rise in DSCs.

Visually, from Figure 5.9, we can see that in MRI, U-Net based multi-modal learning can help the organ segmentation. However, the results of dual-stream networks are not consistent. In some cases, dual-streams V4 improve the segmentation of small organs (kidneys, second row). Meanwhile, in other cases, it is shown that dual-stream V1 gives the best performance in segmenting organs. On the other hand, in CT, multi-modal learning with dual-stream V4 yields consistent results. As shown in Figure 5.10, spleen segmentation is greatly improved by multi-modal learning (first row) and varying organs are better segmented with dual-stream V4 network.

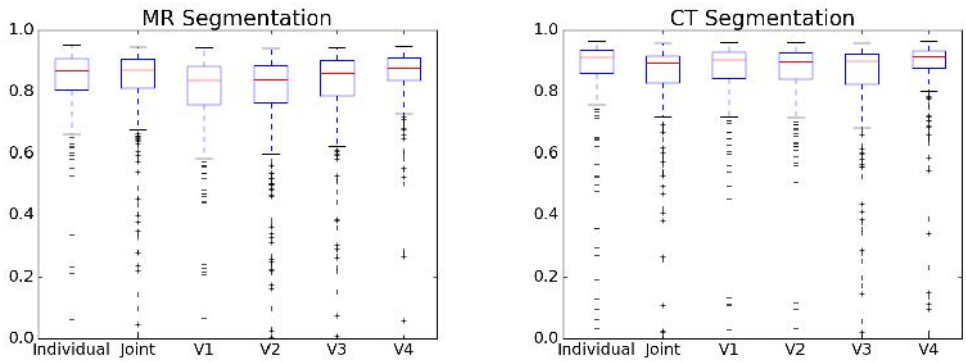


Figure 5.11: Boxplot of MR and CT in FCN based multimodal learning.

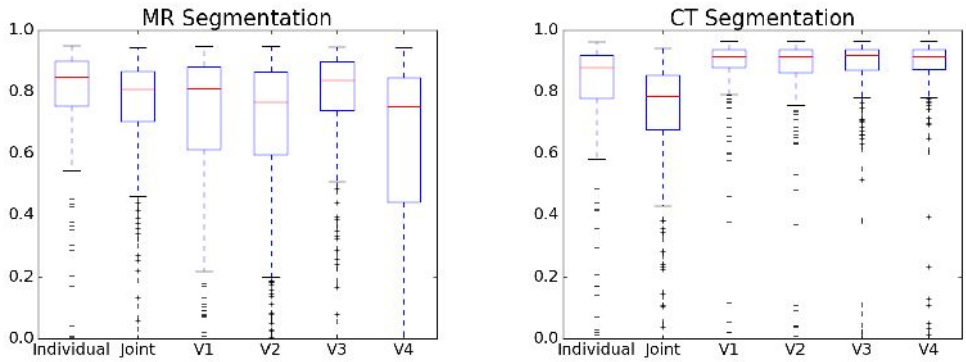


Figure 5.12: Boxplot of MR and CT in UNET based multimodal learning.

5.4 Discussion and Conclusion

In this chapter, we investigate the benefit of multi-modal learning on unpaired CT and MRI images. We present a comparative study of different multi-modal training schemes to better exploit the sharing of information. A novel dual-stream network architecture is introduced for that purpose. By multi-modal learning, shared representation on both modalities can help the network solving the same task with limited training data. Experimental results on both MRI and CT demonstrate improved state-of-the art segmentation accuracies, especially on varying organs such as spleen and kidneys. The power of learning shared representations from different datasets appears as a promising direction for future work.

As we have performed the experiments on multi-modal learning on FCN and U-Net based networks, we have seen different results. Basic architecture differences may give influence on

the behaviour of how the weights between modalities are shared or decoded. Multi-modal learning with U-Net based networks may not be as consistent as FCN-based networks, where dual-stream V4 give the best performance on all of the organs in both modalities (as seen in Fig. 5.11).

Dual-stream networks can increase all of the organ accuracies in CT, compared to individual learning (see Fig. 5.12). However, we noticed that in FCN-based networks, the improvement of multi-modal learning in MRI is not as significant as in CT. The difficulty is due to the inhomogeneous intensity variations within MRI scans. MRI does not have a standardized measurement value per voxel, as it is for example the case for CT. In U-Net based networks, learning with dual-stream networks is even more difficult for MRI, as only two organs can be improved. The difference of our FCN and U-Net based architecture is on the upscore unit in decoder part, where FCN uses addition and U-Net uses concatenation. MRI images are already difficult to train with FCN-based dual-stream networks, compared to CT. In U-Net based networks, it gets more difficult as now the features are concatenated. Hence, there is more complicated latent representation to be interpreted.

In general, for both basic architectures, the average results on a single task (four-organ segmentation) show that multi-modal learning can help to improve the segmentation accuracies on both modalities.

Chapter 6

Learning to Segment Small Organs

This chapter is based on:

Vanya V. Valindria, Ioannis Lavdas, Juan Cerrolaza, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, Ben Glocker. Small Organ Segmentation in Whole-body MRI using a Two-stage FCN and Weighting Schemes. *9th International Conference on Machine Learning in Medical Imaging (MLMI)*, pages 346-354, 2018.

From the multi-organ segmentation results in the previous chapter, there is still left a challenge on small organ segmentation. Hence, in this chapter we study the strategies to segment small structures in whole-body scans. Accurate and robust segmentation of small organs in whole-body MRI is difficult due to anatomical variation and class imbalance. Recent deep network based approaches have demonstrated promising performance on abdominal multi-organ segmentations. However, the performance on small organs is still suboptimal as these occupy only small regions of the whole-body volumes with unclear boundaries and variable shapes.

A coarse-to-fine, hierarchical strategy is a common approach to alleviate this problem, however, this might miss useful contextual information. We propose a two-stage approach with weighting schemes based on auto-context and spatial atlas priors. Our experiments show that the proposed approach can boost the segmentation accuracy of multiple small organs in whole-body MRI scans. We also studied how the method could be applied to another network architecture and compared to end-to-end strategy.

6.1 Introduction

Multi-organ segmentation in whole-body scans is challenging as there are various organs and structures that need to be captured simultaneously. The size, shape and appearance of abdominal organs vary considerably between patients, but also the relative positions change to some degree. Although small organs are rarely investigated before, they are important for abdominal pathology diagnosis, such as cancer screening.

Machine learning methods have been used to segment multiple organs in abdominal images [84]. However, small organs are still underrepresented and show lower accuracies compared to the large ones with less shape variability (e.g., lungs, heart, spine). Small object segmentation is generally more challenging due to large class imbalance between object and background samples. For example, the ratio of small organs in whole-body MRI in our data is less than 0.007% of the overall volume. We focus on bladder, sacrum, rectum, clavicles, pancreas, gallbladder, and adrenal gland. The complexity of background intensity and weak boundaries often make it more difficult to segment small organs.

Most of the multi-organ segmentation works are applied on CT, as the data is widely available. However, MRI gives some advantages over CT, such as lack of radiation and higher detail in soft tissues. Fewer works in MRI organ segmentation is due to the less availability of the datasets and difficulties of the inhomogeneous intensity variations within MR scans. Meanwhile in CT, they have Hounsfield coefficient which is related to specific tissue characterization. MR images only provide relative values and suffer from field inhomogeneities. The fact that most of the multi-organ localization and segmentation are done in CT [134], [32], not all the algorithms are extensible to MRI. The absolute gray values in MR images, unlike CT, are highly dependent on MR imaging parameters.

Therefore, most of the multi-organ segmentation work is applied to CT for which data seems more widely available. We focus on MRI, as there are still fewer works on this modality while whole-body MRI has become an important diagnostic tool for cancer screening. Early works of multi-organ abdominal segmentation include multi-atlas label fusion [7] and statistical

shape models [21]. Multi-atlas techniques register images from a reference database to each new image and fuse multiple atlases to obtain the final segmentation. More recent work has made use of deep learning, for example, architectures for multi-organ segmentation, such as the dense V-network [53], where they argue that the densely linked layers and a shallow V-networks architecture is critical for segmentation accuracy.

Weighted U-Net with weight proportion for foreground and background has been used to address the class imbalance problem [115]. The weight map is precomputed to compensate this imbalance. Unfortunately, the performance of weighted FCN on multi-class segmentation has not been studied yet. Focal loss is introduced by [164] as an attempt for segmentation in class imbalance case. This network outperforms traditional weighted U-Net, however setting the hyper-parameters for a stable training is known to be difficult and tricky - it needs careful initialisation to avoid unstable training.

Two-step networks are currently widely used in both biomedical and general computer vision community to solve small object segmentation. An earlier work that implements this approach is by [116], which combined regional CNN with superpixels at multi-scales. In medical domain, this approach has been applied for lesion segmentation, in pancreas [165] and liver [27], for example. In the first step, FCN is trained to segment the organ as ROI input for a second FCN. The second FCN solely segments lesions from the predicted organ ROIs of previous step. Coarse-to-fine architecture seems to give a much improvement for small object segmentation.

A bounding box with region-of interest (ROI) built on top of predicting segmentation masks inspired by [63], has been implemented by [83] which use regional CNN. Combination of multi-atlas and CNN techniques are used in [83], where they use localization of region of interest using a multi-atlas approach, voxelwise binary classification using a convolutional neural network, and choosing the largest connected component. Coarse-to-fine CNN with region localisation is proposed by [150], which first resizes the image to a smaller scale to scan the whole image to reject most of the background windows, and further refine the segmentation using the cascaded CNN after the object localisation. Coarse-to-fine with combined machine learning approach is then introduced by [161] for kidneys segmentation. First, they localise the organ from whole

body scan using CNN, then segment a kidney on a local image patches with MSL (Marginal Space Learning) and discriminative active shape model. The combination of these methods can reduce the global context failure but still inapplicable for smaller organs than kidneys. A more advanced iterative coarse-to-fine approaches have been done by [166], which use a smaller input region for a more accurate segmentation from multi-view coarse segmentation. Finer segmentation consists of iterative refinement of at least 10 iterations, which can be time consuming.

Pancreas is the most popular small organ to be investigated in previous works, as it is an abdominal organ with very high anatomical variability. Most of the works focus on CT scans and a single organ only, such as the work of [117] that integrates semantic mid-level cues(organ interior and boundary maps) via spatial aggregation that can elevate the performance of segmentation. While [19] proposed recurrent neural network architecture, as a type of long short-term memory (LSTM) to address the contextual learning and segmentation consistency problem. A novel approach with a novel self-attention gating in CNN to segment the pancreas is introduced [106] for a more specific local region segmentation. As an end-to-end approach, they show an improvement in pancreas segmentation, compared to standard FCN, dense dilated FCN [53], holistically nested FCN [117], and standard U-Net, but not outperform the hierarchical/cascade [118] and recurrent CNN [19].

Several works focus on the context module by dilated convolution [154], which is designed to enhance the dense prediction performance by aggregating multi-scale contextual information. To alleviate the missing contextual information in common coarse-to-fine approach, a recurrent saliency transformation network is proposed to relate the coarse and fine stages [155]. This saliency transformation module repeatedly transforms the segmentation probability map from previous iterations as spatial priors in the current iteration. However, they use many iterations, which can be tedious and time consuming. Additionally, some results can be lower than the coarse segmentation results because of the unsatisfying convergence property in iterations. They also do it for small organs, such as gall bladder, adrenal grands, and kidneys, but on CT, not on MRI.

This work is based on a coarse-to-fine framework [19, 154, 166] as it has shown superior in detecting small organs, but we go one step further by incorporating weighting, auto-context, and spatial prior. The advantage of using spatial prior is that coarse-scaled and fine-scaled networks can be jointly optimized.

Our contributions are as follows:
(1) Previous works focus on CT segmentation and a single organ [19, 116, 117, 53, 106], while we study segmentation of multiple small organs on whole-body MRI including structures such as bones, which are rarely considered.
(2) This work is based on a coarse-to-fine framework [19, 154, 166] but goes one step further by incorporating weighting schemes and a specialized ROI selection. Weighting helps with class imbalance. For fine-scale segmentation, we apply auto-context with spatial information obtained from atlases so that coarse-and-fine-scaled networks are jointly optimized[1].

6.2 Methodology

6.2.1 Materials

Whole-body MRI data (MALIBO datasets) are obtained from 48 healthy volunteers using the protocol described in [84]. We segment the same 15 structures in MALIBO as in Chapter 1, with additional small structures, these include 12 abdominal organs (heart, right lung, left lung, liver, adrenal gland, gall bladder, right kidney, left kidney, spleen, pancreas, rectum, and bladder) and 6 bones (spine, right clavicle, left clavicle, pelvis, humerus, sacrum). For our experiments, we only use T2w sequences (because the manual annotations were drawn in T2w images to reduce the misalignment issues), which are resampled to a size (112, 80, 256) with isotropic 4 mm spacing.

[1]Source code available at `https://github.com/vanya2v/Small-organ-segmentation/`

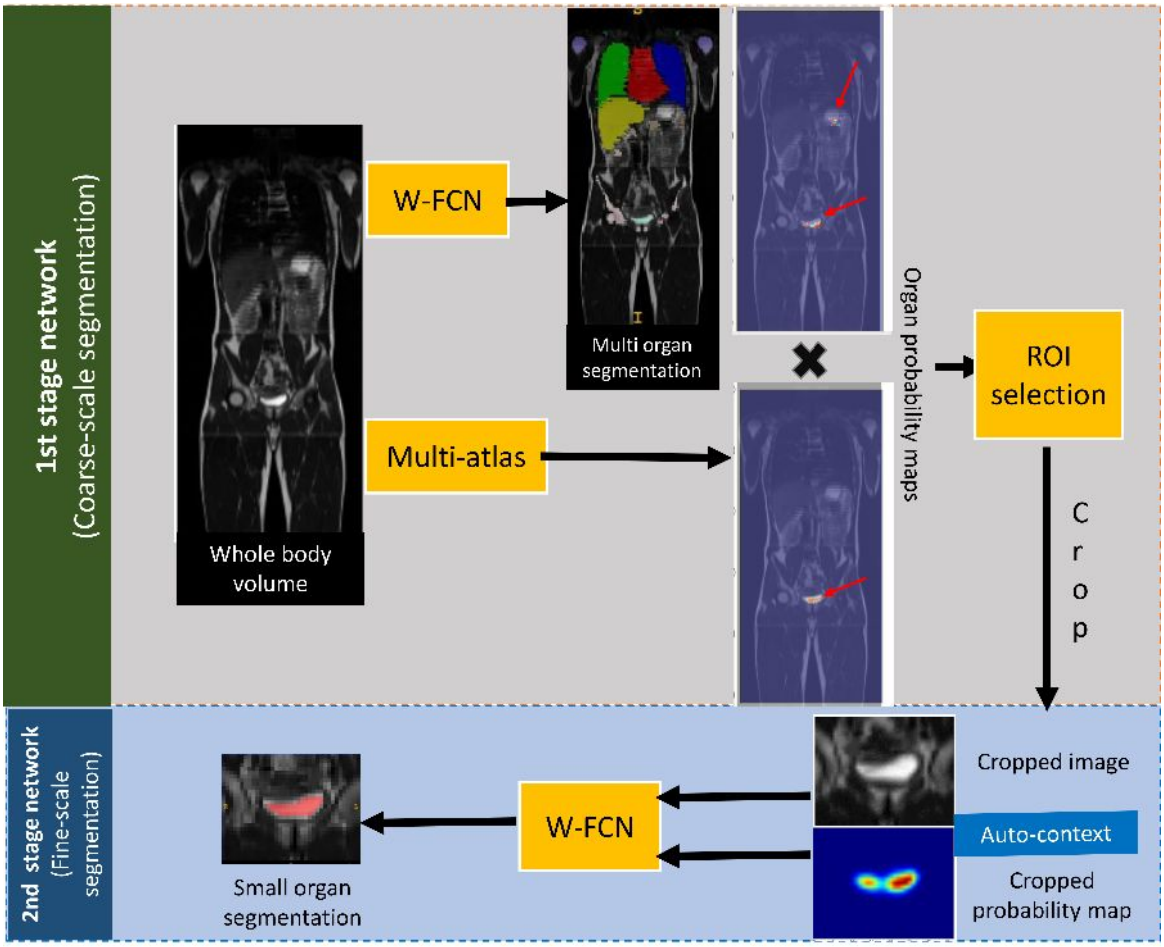


Figure 6.1: Overview. First stage: Coarse-scale segmentation with multi-organ segmentation with weighted-FCN, where we obtain the segmentation results and probability map for each organ. Second stage: Fine-scale binary segmentation per organ. The input consists of a cropped volume and a probability map from coarse segmentation.

6.2.2 Two-stage network: A coarse-to-fine approach

Our aim is to segment multiple small organs from MRI scans, which occupy only a very small part of an MRI volume. We apply a two-stage network, which has been shown to be successful in an organ segmentation task [19, 166, 116]. CNN-based methods produce less accurate results when detecting small organs, particularly because the network is confused by the complicated context in the background and other organs. A coarse-scale is first used to locate the organ of interest for a subsequent fine-scale organ segmentation.

The two-stage strategy effectively reduces the complexity of the background while enhancing the discriminative information of a small organ. We train a coarse-scale segmentation to deal with the multi-organ segmentation on the whole-body scan. Then, at the fine-scale segmentation, we only focus with the ROI selection according to the coarse segmentation. For each stage, we train a separate network. Similarly in testing stage, the coarse-scale network is first used to obtain the rough position of multiple small organs. Then, fine-scale networks are employed for binary segmentation for each of the small organs.

To counter the class imbalance of the small organs, we use a weighted-FCN [115] in both segmentation scales (coarse and fine). For coarse segmentation we apply a weighted-FCN for multi-class segmentation with different weights set for each class. Small weights are set for bigger organs (such as livers, lungs, etc.) and large weights are applied for small organs, according to the relative size to the whole body volumes. This per-class-weight proportion is chosen via experimentation on validation set and the statistics of each organ proportion in the database.

ROI Selection

In whole-body scans, each organ is located in a specific region. Therefore, context and spatial information are crucial for organ segmentation. After organ localisation from the coarse-scale segmentation, we find that in some cases, the FCN fails to locate small organs, leading to a much wider ROI for the fine-scale segmentation. Therefore, we need to induce spatial priors

in order to guide the second stage network with a better ROI. Typically, the segmentation accuracy obtained with multi-atlas based models is lower than the performance of CNN-based frameworks. However, multi-atlas technique is chosen for producing anatomically consistent spatial priors, which can be particularly useful for organ localisation [84].

As shown in Fig.6.1, the bladder segmentation result of multi-atlas is focused in one specific location instead of having multiple predictions scattered in other areas (shown in FCN probability maps).

In the multi-atlas approach, each atlas image is registered to the target and then fused using the methods described in [7]. We chose the PBAF (patch-based segmentation with augmented features) label fusion technique, as compared to other methods, this provides better accuracy for small organs. While [83] used a multi-atlas approach for organ localisation at first step before binary segmentation via CNNs, we only incorporate it after the coarse-scale segmentation. Here, the probability map of coarse-segmentation is multiplied with the one from the multi-atlas approach (as a spatial prior for false-positive reduction, so that the fidelity of the final probability map is improved) to produce the 3D bounding box for fine-scale segmentation, see Fig.6.1 top for details. We denote this part of spatial prior incorporation as *combined multi-atlas*.

Auto-context

We utilize the classical approach of auto-context [136] into our framework to fuse and to integrate the information from different stages with the context. Auto-context has been applied in high-level vision for an effective context model learning. This method integrates image appearance with the context information (classification confidence) by learning a series of classifiers [136]. Zografos [172] uses a hierarchical auto-context for multiple organ segmentation in CT. They extract rich features from multiscale supervoxels, 3D GLCM, shape context features, HOG-3D features, and neighborhood context to be fed to the hierarchical auto-context classification. To determine the best labeling, they apply Conditional Random Field (CRF) step to merge the auto-context output.

In our cases, the problem of coarse-to-fine segmentation is that sometimes the cropped ROI within the bounding box has less sufficient spatial context, making the fine-scale networks more confused than the coarse-scale segmentation [117]. Hence, we incorporate the probability maps from the coarse-scale segmentation into the fine-scale segmentation. The benefit of probability maps as visual cues have been discussed in [117] as a spatial aggregation from multi-view segmentation and in [155] as an updated input for an iterative fine-scale segmentation with saliency transformation network. In our case, a simple auto-context with probability maps incorporation helps the FCN to integrate the information from coarse-to-fine level segmentation. By using additional input (see Fig.6.1 bottom), uncertainties and errors from the first stage are adjusted.

6.3 Experiments and Results

6.3.1 FCN-based Segmentation

We use different strategies to achieve small organ segmentation. For coarse and fine scale segmentation, we use the same architecture of an FCN with residual layers, as implemented in [110]. We train the network using Adam optimisation with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 10^{-5}$. In coarse-scale segmentation, we use mini-batch training of 16 training examples with size 64^3, to provide enough context at 4 mm resolution. Meanwhile in fine-scale segmentation, the example size are 8^3 to fit with the smaller input image size. We train each networks with 10K iterations.

We perform two-fold cross-validation, where we split the dataset into two fixed folds with equal number of samples. For evaluation, we measure the segmentation accuracy by computing the Dice Similarity Coefficient (DSC) for each subject.

First stage of the network is trained for multi-organ (18 classes, including large organs) on the entire 3D whole-body ROI. We can see different results from different techniques applied in Table 6.1. Although, the overall results for multi-atlas segmentation are lower than baseline

FCN, but the accuracies for small organs are better. For the baseline FCN, the small organs segmentation still give unsatisfying accuracies.

Then, we investigate the role of introducing different weights in taking training samples from different classes on coarse-scale multi-organ segmentation. As we give larger weights on small organs, we can reduce the effect of class imbalance [115]. This weighted FCN gives impressive increment of small organ accuracies compared to the baseline in Table 6.1, with overall small organ accuracies increase from 0.244 to 0.483 Dice scores.

However, for our study, we are only interested in small organs as large organs, such as heart, liver, lungs, and spine - have already achieved satisfactory results (with above 0.9 DSC). Our complete benchmarks for different strategies applied in small organ segmentation are shown in Table 6.2. Although the overall results on 18 organs for multi-atlas segmentation are lower (DSC: 0.486) than baseline FCN (DSC: 0.516) (without weights, auto-context, two-stage network and combined multi-atlas), the accuracy for small organs improved.

To evaluate the spatial prior in combined multi-atlas, we multiply the probability maps from the weighted-FCN with the multi-atlas prediction. Table 6.2 shows that it slightly improves the accuracy for small organ segmentation, especially adrenal gland (DSC: 0.203). However, some organs, such as both clavicles are worse because of the lower accuracy on multi-atlas segmentation. The probability maps given by multi-atlas are tighter, so that the segmentation misses some parts of the region of interested.

To verify that the two-stage network can segment small organs more accurately, we run the state-of-the-art method [165], which takes input from baseline FCN segmentation to crop the ROI for fine-scale segmentation. Compared to the small organ segmentation accuracies produced by the FCN baseline (DSC: 0.244), we observe an improvement of about 50%. This result shows the advantage of using a two-stage network on small organ segmentation.

We then employ a two-stage scheme using the weighted-FCN with auto-context. We use the probability map from weighted-FCN prediction as an additional input to the network. With this strategy, we get a much higher accuracy on all organs (see Table 6.2). Weightings, auto-

Table 6.1: DSCs of coarse multi-organ segmentation by multi-atlas, baseline FCN , weighted FCN (W-FCN) and weighted-FCN with spatial prior (W-FCN-S). Small organs are shown in italics

ID	Organ	Multi-atlas [7]	FCN	W-FCN [115]	W-FCN-S
1	Heart	0.193	0.847	0.898	0.898
2	Rlung	0.767	0.928	0.941	0.941
3	Llung	0.765	0.924	0.943	0.943
4	Liver	0.728	0.873	0.900	0.900
5	*Adrenal Gland*	*0.044*	*0.006*	*0.097*	***0.203***
6	*Gall Bladder*	*0.298*	*0.330*	***0.443***	*0.436*
7	Rkidney	0.710	0.692	0.784	0.784
8	Lkidney	0.634	0.648	0.745	0.745
9	Spleen	0.503	0.613	0.685	0.685
10	*Pancreas*	*0.377*	*0.373*	*0.465*	***0.519***
11	*Bladder*	*0.680*	*0.401*	*0.689*	***0.717***
12	Spine	0.749	0.780	0.867	0.867
13	*R. Clavicle*	*0.135*	*0.055*	***0.462***	*0.381*
14	*L. Clavicle*	*0.163*	*0.080*	***0.484***	*0.397*
15	Pelvis	0.591	0.325	0.759	0.759
16	Humerus	0.525	0.706	0.831	0.831
17	*Rectum*	*0.429*	*0.346*	*0.526*	***0.549***
18	*Sacrum*	*0.466*	*0.361*	***0.701***	*0.693*
	Avg	0.486	0.516	0.679	**0.680**
	StdDev	0.238	0.303	0.226	**0.219**
	Small org avg	0.324	0.244	0.483	**0.487**

Table 6.2: Different strategies on segmentation of small organs, using the baseline FCN, weighted-FCN (W), auto-context (AC), two-stage networks (2SN), and combined multi-atlas (CMA) for ROI selection. Improvement in overall small organ segmentation accuracy (reported in DSCs) is achieved with our proposed approach.

Approaches				Adrenal Gland	Gall Bladder	Pancreas	Bladder	R. Clavicle	L. Clavicle	Rectum	Sacrum	Avg.
Multi-atlas				0.044	0.298	0.377	0.680	0.135	0.163	0.429	0.466	0.324
FCN												
W	AC	2SN	CMA									
✗	✗	✗	✗	0.006	0.330	0.373	0.401	0.055	0.080	0.346	0.361	0.244
✓	✗	✗	✗	0.097	0.443	0.465	0.689	0.462	0.484	0.526	0.701	0.483
✓	✗	✗	✓	**0.203**	0.436	0.519	0.717	0.381	0.397	0.549	0.693	0.487
✗	✗	✓	✗	0.046	0.375	0.455	0.594	0.274	0.379	0.393	0.490	0.376
✓	✓	✓	✗	0.133	0.507	**0.612**	0.729	0.519	0.535	0.591	0.721	0.543
✓	✓	✓	✓	0.146	**0.532**	0.567	**0.754**	**0.541**	**0.547**	**0.658**	**0.735**	**0.560**

Table 6.3: Paired t-tests are performed for significance analysis against **FCN** as the baseline, with the significance threshold at p= 0.05

Strategy	Mean DSC	Variance	p-value
FCN	0.244	0.063	*
FCN+W	0.483	0.058	7.46E-62
FCN+2SN	0.376	0.073	7.22E-28
FCN+2SN+W+AC+CMA	0.560	0.056	4.54E-74

context, and region cropping are shown to boost the performance of small organ segmentation.

To add the spatial prior, we crop the ROI for fine-scale segmentation for organ-specific bounding box, according to the Section 6.2.2. As detailed in Table 6.2, adding spatial priors is useful in almost all small organs, except the pancreas because of its poor tissue contrast and shape variability. For adrenal gland, too, we find that the results of direct multiplication between coarse-scale segmentation and spatial prior gives better results. Multi-atlas prior information gives better localisation for adrenal gland, which is the smallest organ in our task (only occupied about 0.0001% of the whole volume). Overall, the addition of auto-context and spatial priors to two-stage weighted-FCN gives the best results on small organ segmentation (DSC: 0.560), as detailed in Table 6.2.

Paired two-tailed t-tests are performed for significance analysis against the chosen baseline method. We used a paired Students t-test to determine whether the performance difference between two segmentation strategies was significant. In Table 6.3, we show that all of the additional strategies (weighting, two-stage network, auto-context, and spatial prior incorporation) significantly outperformed the baseline FCN. In addition, as seen in the Table 6.3, weighting gives more significant improvement to the performance, compared to two-stage network (as the smaller the p-value, the stronger the evidence is that the two strategies have different means). Table 6.4 shows paired t-tests for significance analysis against two-stage FCN. From the table, we can see that weighting and auto-context significantly improve the performance of baseline two-stage FCN. Incorporation of spatial prior from multi-atlas also makes the observed segmentation performance difference, with $p < 0.05$.

Table 6.4: Paired t-tests are performed for significance analysis against **Two-stage FCN** as the baseline, with the significance threshold at p= 0.05

Strategy	Mean DSC	Variance	p-value
FCN+2SN	0.376	0.073	*
FCN+2SN+W+AC	0.543	0.061	2.25E-41
FCN+2SN+W+AC+CMA	0.560	0.056	0.009

6.3.2 DeepMedic based Segmentation

To further evaluate our proposed approach, we also experiment with another network architecture, which is DeepMedic (DM) [74]. The difference from FCN architecture is DeepMedic uses multi-scale 3D CNN coupled with a 3D fully connected Conditional Random Field (CRF), which is specifically designed for medical image segmentation [74].

We apply the same strategies done in the previous subsection with FCN-based architecture. First stage of the network is trained for multi-organ (18 classes, including large organs) on the entire 3D whole-body ROI. We can see different results from different techniques applied in Table 6.5. Compared to the baseline FCN (DSC: 0.516), baseline DeepMedic achieved better accuracies (DSC: 0.616) as it is specifically designed for medical image segmentation. The class weighting schemes in DeepMedic also gives a better accuracy, instead of using a uniform weight samples between background and foreground. The weights used per organ are the same as the ones used in FCN-based networks (based on organ size proportion). Table 6.5 shows that combining spatial prior with weighted DeepMedic slightly decreases the accuracy for small organ segmentation, but gives the best accuracy for adrenal gland (DSC: 0.247). However, adding spatial prior may improve the segmentation in half of the small organs.

For the two-stage network scheme, we take the cropped ROI from the baseline DeepMedic segmentation for fine-scale segmentation. We observe an improvement for almost 50% compared to the small organ segmentation accuracies produced by the DeepMedic baseline (DSC: 0.321). This result shows the advantage of using a two-stage network on small organ segmentation. We then add auto-context to the two-stage network of weighted-DM, which gives a better accuracy (DSC: 0.562).

Table 6.5: DSCs of coarse multi-organ segmentation by multi-atlas, baseline DeepMedic, weighted DM (W-DM) and weighted-DM with spatial prior (W-DM-S). Small organs are shown in italics

ID	ORGAN	Multi-atlas	DM[74]	W-DM	W-DM-S
1	Heart	0.193	0.891	0.888	0.888
2	Rlung	0.767	0.941	0.928	0.928
3	Llung	0.765	0.937	0.928	0.928
4	Liver	0.728	0.911	0.906	0.906
5	Adrenal Gland	0.044	0.000	0.164	**0.247**
6	Gall Bladder	0.298	0.004	0.499	**0.517**
7	Rkidney	0.710	0.799	0.854	0.854
8	Lkidney	0.634	0.742	0.820	0.820
9	Spleen	0.503	0.740	0.707	0.707
10	Pancreas	0.377	0.506	0.538	**0.583**
11	Bladder	0.680	0.683	0.670	**0.761**
12	Spine	0.749	0.875	0.861	0.861
13	R. Clavicle	0.135	0.013	**0.559**	0.443
14	L. Clavicle	0.163	0.045	**0.587**	0.457
15	Pelvis	0.591	0.819	0.779	0.779
16	Humerus	0.525	0.857	0.853	0.853
17	Rectum	0.429	0.626	**0.639**	0.632
18	Sacrum	0.466	0.692	**0.742**	0.721
	Avg	0.486	0.616	**0.718**	0.716
	Small org avg	0.324	0.321	0.550	0.545

Table 6.6: Different strategies on segmentation of small organs, using the baseline DM, weighted-DM (W), auto-context (AC), two-stage networks (2SN), and combined multi-atlas (CMA) for ROI selection. Improvement in overall small organ segmentation accuracy (reported in DSCs) is achieved with our proposed approach.

Approaches				Adrenal Gland	Gall Bladder	Pancreas	Bladder	R. Clavicle	L. Clavicle	Rectum	Sacrum	Avg.
Multi-atlas				0.044	0.298	0.377	0.680	0.135	0.163	0.429	0.466	0.324
FCN												
W	AC	2SN	CMA									
✗	✗	✗	✗	0.000	0.004	0.506	0.683	0.013	0.045	0.626	0.692	0.321
✓	✗	✗	✗	0.164	0.499	0.538	0.670	0.559	0.587	0.639	0.742	0.550
✓	✗	✗	✓	0.247	0.517	0.583	0.761	0.443	0.457	0.632	0.721	0.545
✗	✗	✓	✗	0.000	0.480	0.547	0.761	0.169	0.342	0.621	0.738	0.457
✓	✓	✓	✗	0.163	0.527	0.549	0.698	0.560	0.612	0.654	0.734	0.562
✓	✓	✓	✓	0.207	0.547	0.567	0.811	0.600	0.607	0.679	0.732	**0.594**

Lastly, similar to the previous section, we employ a two-stage scheme using the weighted-DM with auto-context and the probability map from the multi-atlas prediction as an additional input to the network. With this strategy, we get the best overall accuracies (see Table 6.6). Weightings, auto-context, and region cropping by spatial prior incorporation are shown to boost the performance of small organ segmentation in DeepMedic based architecture.

6.3.3 Comparison to end-to-end Strategies

In the previous subsection, we proposed an approach to segment small organs based on two-stage networks, which is not an end-to-end approach. Although limited, end-to-end solutions for multi-organ segmentation exist, as proposed by [106, 112, 53].

In this section, we try to segment small organs using autofocus layer networks proposed by [112] with DeepMedic-like multi-scale networks. Autofocus layer is a module to choose the optimal scale of detecting different objects in an image. The attention mechanism [6] is used to indicate the importance of each scale when processing different locations in an image.

An interesting principle of autofocus is its capability to 'zoom in or zoom out' to use different context. We also compare the autofocus network (AFN) idea with another approach, such as dilated convolution [154].D DeepMedic uses different pathways with high and low resolution inputs. In BASIC networks, we only keep DeepMedic's high resolution pathway and dilate the last 6 hidden layers (rate 2) to enlarge the receptive field. In ASPP (Atrous Spatial Pyramid Pooling) proposed by [25], dilated convolution can process greater context without

Table 6.7: End-to-end strategies on segmentation of small organs. Organ segmentation Dice score of with DeepMedic (DM) as the baseline, compared to BASIC network, with dilated-convolutions (ASPPc and ASPPs) and autofocus layers (AFN)

ID	ORGAN	Basic	ASPP_c	ASPP_s	DM	AFN1	AFN2	AFN3	AFN4	AFN5	AFN6
1	Heart	0.8952	0.9001	0.8964	0.8911	0.9049	0.8831	0.8989	0.9048	0.8959	0.8990
2	Rlung	0.8580	0.9238	0.8951	0.9412	0.9285	0.9103	0.9448	0.9447	0.9460	0.9466
3	Llung	0.8623	0.9242	0.9055	0.9374	0.9332	0.9215	0.9431	0.9433	0.9450	0.9435
4	Liver	0.9026	0.9036	0.8966	0.9113	0.9139	0.8727	0.9141	0.9150	0.9124	0.9116
5	Adrenal Gland	0.0000	0.0023	0.0042	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	Gall Bladder	0.3049	0.2836	0.2569	0.0035	0.3703	0.1826	0.4251	0.4094	0.3925	0.4389
7	Rkidney	0.7073	0.4855	0.5181	0.7990	0.7803	0.5748	0.7626	0.7908	0.7587	0.7816
8	Lkidney	0.6564	0.5802	0.6328	0.7417	0.7218	0.5017	0.7248	0.7401	0.6926	0.7266
9	Spleen	0.6852	0.6457	0.6553	0.7396	0.6962	0.6425	0.6932	0.7094	0.6795	0.6773
10	Pancreas	0.4643	0.4891	0.4602	0.5062	0.5410	0.4384	0.5690	0.5496	0.5524	0.5648
11	Bladder	0.5602	0.6034	0.6479	0.6833	0.6508	0.5928	0.6256	0.7227	0.6870	0.6680
12	Spine	0.8414	0.8357	0.8405	0.8749	0.8385	0.8249	0.8393	0.8425	0.8380	0.8342
13	R. Clavicle	0.1041	0.2320	0.2306	0.0129	0.2313	0.1145	0.2649	0.2862	0.2728	0.2732
14	L. Clavicle	0.1142	0.2723	0.2647	0.0449	0.2557	0.0672	0.2803	0.2815	0.3148	0.3001
15	Pelvis	0.6403	0.6947	0.6573	0.8190	0.6472	0.6274	0.6413	0.6753	0.6852	0.6668
16	Humerus	0.2786	0.3282	0.2964	0.8572	0.2919	0.1928	0.3409	0.3334	0.4044	0.3585
17	Rectum	0.1635	0.4654	0.5065	0.6264	0.5037	0.3142	0.4328	0.5737	0.5215	0.5127
18	Sacrum	0.5541	0.6372	0.5774	0.6917	0.5871	0.5044	0.6079	0.6036	0.5958	0.5688
	Avg	0.5329	0.5670	0.5635	0.6156	0.5998	0.5092	0.6060	**0.6237**	0.6164	0.6151
	Small org avg	0.2832	0.3731	0.3685	0.3211	0.3925	0.2768	0.4007	**0.4283**	0.4171	0.4158

downsampling the feature maps. ASPP-c denotes the fusion of activation from all scales via concatenation, while ASPP-s via summation. AFN networks are obtained by converting the last n hidden layers of BASIC to autofocus layers, denoted as "AFN-n", where $n = \{1, ..., 6\}$.

Our results with the original DeepMedic architecture are shown in Table 6.7. All of the strategies use the uniform sampling between background and foreground. From the Table 6.7, BASIC dilated convolution cannot enhance the overall and small organ segmentation, it is worse than our baseline DeepMedic (DM), which uses both low and high resolution pathway. But ASPP-c and ASPP-s gives an improvement to the DM segmentation accuracies, especially on small organs.

As expected, better results are achieved with AFN networks in overall and small organ accuracies. It is because parallel dilated convolutional filters can capture both local and more global context. Additionally, from our results in Table 6.7, AFN-4 with last 4-layers autofocus layers gives the best small-organ accuracy (DSC: 0.4283). Most of the small organs are improved with AFN networks, except rectum and sacrum.

Visual results are shown in end-to-end examples of Figure 6.2. The results are shown as multiple

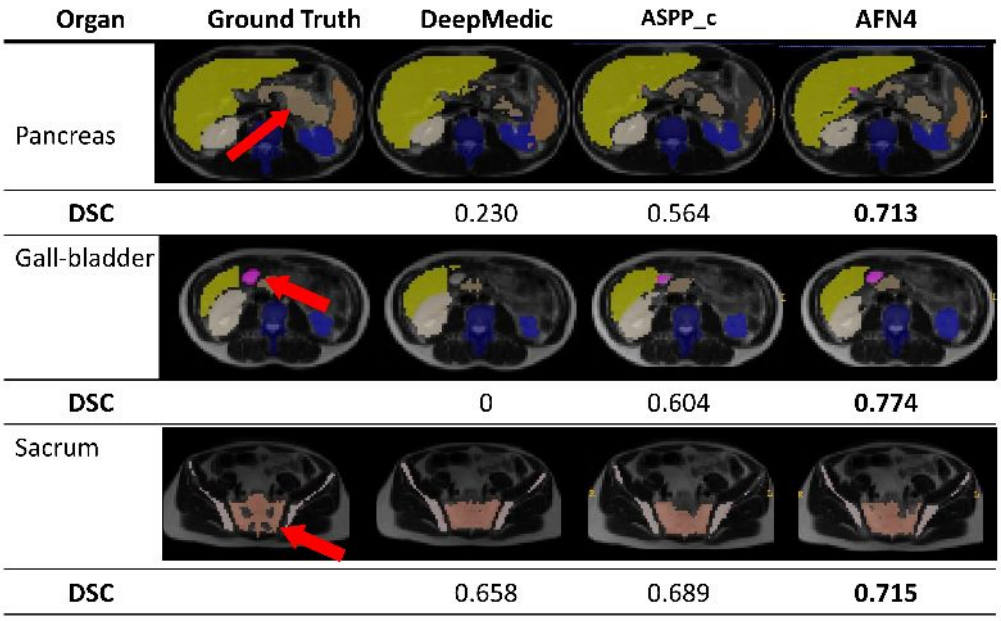


Figure 6.2: Fine-scale segmentation of small organs (pancreas, gall-bladder and sacrum) with end-to-end AFN

classes with multi-organ segmentation. We choose ASPP-c as it gives the best result for dilated convolution and AFN4 as it yields the best overall results. Pancreas and sacrum segmentation in Figure 6.2 are improved with AFN4. With baseline DeepMedic, gall-bladder cannot be detected (Figure 6.2 middle), but the use of dilated convolution in ASPP-c can help the organ detection, which is further improved by AFN4.

Additionally, we evaluate the weighting schemes with the same weights in the previous experiments per class to end-to-end strategies. We can see in Table 6.8, the accuracy of AFN weighted segmentation are mostly higher than the standard AFN (Table 6.7). However, compare to the baseline (DM-w: Deep Medic weighted), the accuracy is still lower. This is due to the nature of the auto focus layer, which selects a scale according to a structure's size automatically and does not depend heavily on the weights.

6.4 Discussion and Conclusion

Small organ segmentation is a very challenging task, even for humans with specialized radiological and anatomical background. The difficulty in automatic small organ segmentation is

Table 6.8: End-to-end strategies on segmentation of small organs. Organ segmentation Dice score of with DeepMedic weighted (DM-w) as the baseline, compared to BASIC network, with dilated-convolutions (ASPPc and ASPPs) and autofocus layers (AFN)

ID	ORGAN	Basic	ASPP_c	ASPP_s	DM-w	AFN1	AFN2	AFN3	AFN4	AFN5	AFN6
1	Heart	0.8876	0.8872	0.8931	0.8884	0.8871	0.8906	0.8994	0.8937	0.8976	0.8946
2	Rlung	0.8311	0.9226	0.9217	0.9285	0.9169	0.9335	0.9344	0.9442	0.9474	0.9434
3	Llung	0.8198	0.9197	0.9153	0.9281	0.9144	0.9282	0.9333	0.9413	0.9449	0.9419
4	Liver	0.9084	0.9114	0.9108	0.9064	0.9106	0.9142	0.9110	0.9161	0.9150	0.9115
5	Adrenal Gland	0.0000	0.0000	0.0023	0.1635	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	Gall Bladder	0.4294	0.3820	0.4410	0.4991	0.4168	0.4519	0.4817	0.4730	0.4623	0.4471
7	Rkidney	0.7359	0.7851	0.7633	0.8541	0.7737	0.7568	0.7758	0.7812	0.7819	0.7891
8	Lkidney	0.6926	0.7321	0.7172	0.8202	0.7228	0.7213	0.7199	0.7284	0.7231	0.7284
9	Spleen	0.7160	0.6513	0.6708	0.7074	0.6882	0.6451	0.6738	0.6865	0.6898	0.6924
10	Pancreas	0.5111	0.5151	0.5117	0.5385	0.5516	0.5254	0.5402	0.5815	0.5911	0.5657
11	Bladder	0.7108	0.7007	0.7291	0.6704	0.7272	0.7188	0.7604	0.7698	0.7530	0.7530
12	Spine	0.8439	0.8381	0.8399	0.8609	0.8398	0.8333	0.8425	0.8470	0.8457	0.8439
13	R. Clavicle	0.1109	0.1677	0.2151	0.5587	0.1480	0.1045	0.1509	0.1904	0.1950	0.1077
14	L. Clavicle	0.0194	0.1876	0.2485	0.5871	0.1566	0.0779	0.1672	0.1644	0.1708	0.2308
15	Pelvis	0.7141	0.6561	0.6790	0.7790	0.6827	0.6820	0.6879	0.7193	0.7099	0.7054
16	Humerus	0.4181	0.4009	0.3893	0.8528	0.4076	0.4060	0.4194	0.4253	0.4304	0.4290
17	Rectum	0.5315	0.5806	0.5742	0.6390	0.6083	0.6132	0.6493	0.6130	0.6136	0.5716
18	Sacrum	0.6405	0.6349	0.6269	0.7418	0.6534	0.6244	0.6462	0.6650	0.6674	0.6477
	Avg	0.5845	0.6041	0.6138	**0.718**	0.6114	0.6015	0.6219	0.6300	0.6299	0.6224
	Small org avg	0.3692	0.3961	0.4186	**0.5498**	0.4077	0.3895	0.4245	0.4321	0.4317	0.4155

the main motivation of this work. A standard CNN-based approach gives good performance on larger organs, but still fails to accurately segment small organs. We found that by setting different weights on training examples per class could boost the performance of small organ segmentation. The class imbalance problem is countered by a multi-class weighted-FCN.

As the target is often very small, it is also required to focus on a local input region. The two-stage network scheme seems to help the small organ segmentation. However, lack of contextual information and spatial knowledge make the network confused. Hence, we apply a simple auto-context, which use the coarse-scaled probability map to carry useful context information for fine-scale segmentation. Adding spatial prior from a more traditional multi-atlas segmentation approach is also shown to help focusing the organ-specific region for fine-scale segmentation.

Figure 6.3 show the cases where a two-stage network without auto-context and spatial priors fails. Using only a two-stage network, it fails to localise the fine-scale ROI, resulting in a worse segmentation result. Auto-context could help to select the correct ROI. Adding spatial priors from a multi-atlas segmentation approach is also shown helpful to identify the organ-specific region for fine-scale segmentation. Some examples of small organ fine-scaled segmentation

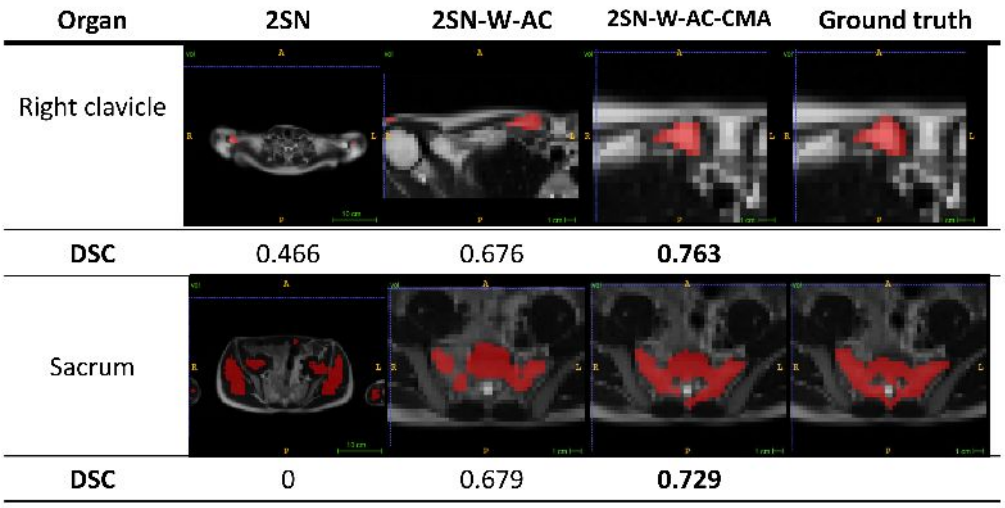


Figure 6.3: Two-stage network (2SN) without weighting (W) and auto-context (AC) miss the local ROI selection on some organs (right clavicles and rectum), while adding spatial prior from multi-atlas (CMA) could enhance the segmentation results.

are shown in Figure 6.4. We can see that the two-stage weighted network with auto-context and combination with multi-atlas produce better results in small organ segmentation. False positives are reduced and the ROI is better focused on the organ.

Similar cases are also shown with DeepMedic based segmentation. Figure 6.5 illustrate the superiority of adding spatial prior to the weighted two-stage network. A clear example can be seen in gall-bladder case, where using only two-stage network fails to detect the small organ completely (Fig.6.5 middle) and adding auto-context with spatial prior can improve the segmentation accuracy. Examples of cases where ROI selection has influence on segmentation accuracies are depicted in Figure 6.6. A basic two-stage network capture a bigger ROI on right clavicle, hence misses most of the organ. With additional auto-context and spatial-prior, the segmentation accuracy of right clavicle is greatly improved. The same case is shown in sacrum segmentation.

End-to-end approach (dilated convolution and AFN networks) is indeed less time-consuming, less-complicate than weighted-spatial two-stage network. However, the improvement of small organs segmentation is also less (DSC: 0.428) than our proposed approach (DSC: 0.594), in DeepMedic based architecture. While two-stage networks gives improvement in all of the organs

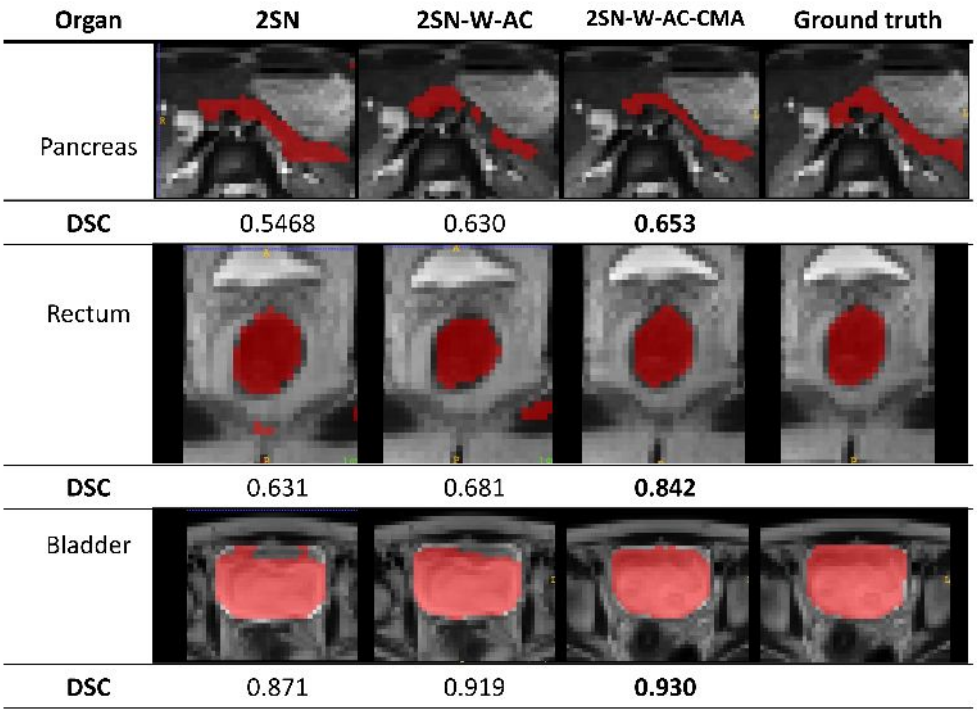


Figure 6.4: Fine-scale segmentation of small organs (pancreas, recum, and bladder). Two-stage network (2SN) with weighting (W), auto-context (AC), and combined spatial prior from multi-atlas (CMA) can improve small organ segmentation results.

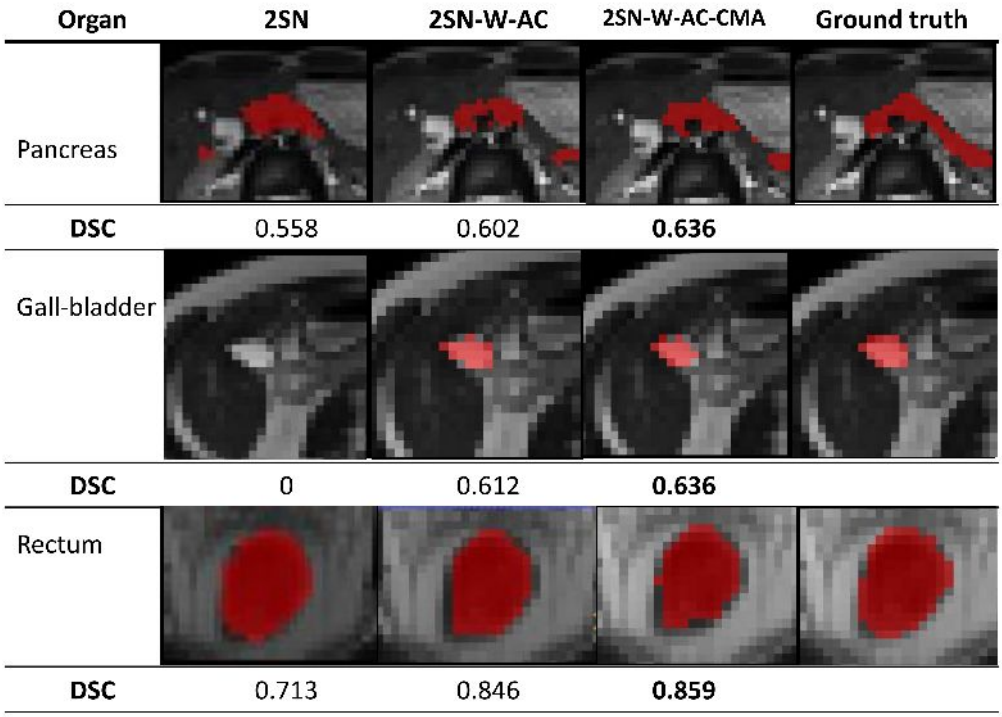


Figure 6.5: Fine-scale segmentation of small organs (pancreas, gall-bladder, rectum) in DeepMedic based segmentation. Two-stage network (2SN) with weighting (W), auto-context (AC), and combined spatial prior from multi-atlas (CMA) can improve small organ segmentation results.

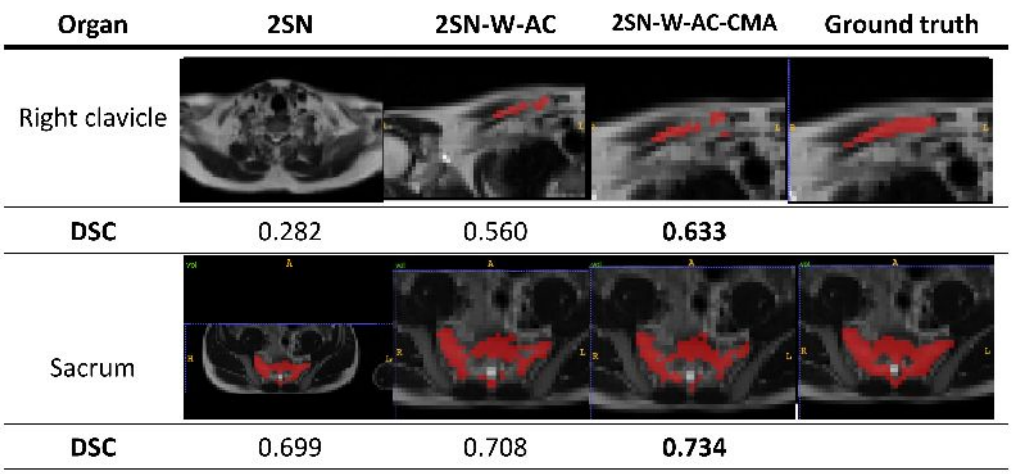


Figure 6.6: Two-stage network (2SN) without weighting (W) and auto-context (AC) miss the local ROI selection with larger FOV on some organs (right clavicles and sacrum), while adding spatial prior from multi-atlas (CMA) could enhance the segmentation results.

than basic networks, all of the end-to-end AFNs give worse accuracies in two organs (rectum and sacrum). This is due to a more complicated and 'hard-to-control' process of autofocus, which select different context or organ automatically. Moreover, in all of the end-to-end strategies, adrenal gland cannot be detected at all. But, our two-stage network can show some adrenal gland segmentation, although only with small improvement (around 0.2 Dice scores). Weighting schemes in autofocus networks also do not help as much as in two-stage networks, as the main task of AFN is to automatically choose the correct focus to a specific structure.

Our two-stage approach with weighting and spatial prior is applied to multiple small organ segmentation, which outperforms the baseline (standard FCN and previous state-of-the art) significantly. This work shows a promising result for small organs and bone segmentation in whole-body MRI.

Chapter 7

Summary and Future Work

In this chapter, we first give a brief summary of the thesis and the research problems. In addition, we discuss about the current limitations of the proposed algorithms and list possible directions for future research. Lastly, we conclude the chapter with a general outlook on the thesis.

7.1 Summary

The use of whole-body imaging techniques has been very useful for abdominal disease diagnosis, including bronchial carcinoma, hepatic malignancies, renal carcinoma, colonic cancer, lymphoma, and also rare malignancies such as bone or soft tissue tumors [138]. Whole-body MRI are complex scans containing a lot of information compared to more focused scans (e.g., a scan of the brain). Whole-body MRI is able to cover total body screening from head to toes in one diagnostic work-flow. MRI allows accurate assessment of whole-body to generate an unprecedented snapshot of patient health with no side effects. Moreover, image based quantitative analysis has been increasingly recognised as important for more detailed and objective analysis of the human anatomy. This includes derivation of quantitative, statistical (e.g., organ volumetric measurement), and qualitative (e.g., lesion detection) results from the imaging data.

Image analysis tools have an important role in accurate and objective assessment of these

clinical measurements from imaging data. However, full automation and reliable utilisation of the analysis tools still face technical challenges due to limitations imposed by the imaging modalities and practical issues:

1. Imaging data quality. As the scans are acquired in subsequent stations, scans with repeated station or missing stations might be present. Problems with imaging protocols in whole-body MRI also introduce fat-water swap (see Chapter 2) in scans, which brings difficulties in further automatic image analysis.

2. Few labeled data. Whole-body MRI covers many anatomical structures, with more than 10 important organs and bones. With these numbers, it would be time consuming for an expert to label all of the structures from large-number of patients. Hence, our limitation in developing automatic tools for whole-body images is lack of labeled data.

3. Different imaging modalities. In clinical practice, it is common to switch the imaging modality from MRI to CT, and vice versa, dependent of the diagnosis and the availability of the machine in the hospital. However, different scans from different modalities impose differences in image appearance, which can be a real challenge in deploying automatic image analysis tool.

The thesis has focused on studying the effects of these limitations imposed by the imaging modalities on common automated analysis tasks such as image segmentation, quality control and domain adaptation. The experimental results, provided in the previous chapters, show that our proposed frameworks can lead to more reliable solutions.

Our presented approaches bring up common features:

a. The use of recent machine learning models for better understanding of image data

b. New image representation to provide better segmentation accuracy

The use of machine learning brings an automated image analysis framework, which is later used to guide the analysis and to reduce the manual human work. For instance, the RCA proposed in Chapter 3 can inform and guide whether the automated segmentation results have good or

bad accuracies, without the need of manual delineation. We also assess the application of RCA to solve a common clinical problem with domain adaptation between datasets from different imaging centers. In Chapter 4, we show that selecting the 'most useful' samples with RCA from target domain can save manual annotation time with better results obtained through supervised transfer learning.

Learning new image representations, on the other hand, offers better data interpretability in image analysis. For example, we can consider the multi-modality experiments in Chapter 5, where MR and CT abdominal images are learned together in the latent space. By sharing or merging the modality-specific features in dual-stream networks, we can extract useful properties of each modality for multi-organ segmentation. The learned latent multi-modal space enables us to perform multi-organ segmentation in unpaired MR and CT images using a single dual-stream network. We also evaluate the multi-modal learning with different basic architectures.

Strategies to alleviate the difficulty of segmenting small organs are explained in Chapter 6. The use of two-stage networks with weighting and spatial prior addition is shown to be able to capture smaller structures of whole-body scans. The local ROI selection enables us to perform a finer segmentation, where the background intensity is less complicated. In this way, the segmentation task is simplified. In addition, we compare our strategies with different basic network architectures and end-to-end strategies.

Taking the experimental results into consideration, we also review the strengths and limitations of the various methods over each other, with the aim of providing a guideline for choosing the most suitable automated tool for future study. We conclude that machine learning techniques are very useful and important to solve different challenges in whole-body image analysis. This will bring benefit for automatic image analysis in a larger population study. However, some limitations will be discussed in the following section.

7.2 Limitations and Future Work

In this section, the main limitations of the presented image analysis methods are listed together with some suggestions to tackle these limitations in future research.

Learning to predict segmentation accuracy

RCA works very well for DSC prediction, as DSC is the common measure for evaluating the segmentation accuracy. However, RCA prediction is still less accurate for other evaluation metrics, such as surface-based distance metrics. This limitation is not that crucial as commonly we only assess the DSC to identify segmentation quality after deployment of a segmentation methods.

Our next goal is to apply RCA as an effective tool to extract the subset of high quality segmentation in large-scale imaging studies, such as the UK Biobank. We believe that RCA will be useful as an automatic quality control tool where thousands of images are automatically analysed.

Learning between different domains

Then we describe an application of RCA in clinical practice to solve domain adaptation problem. Although the approach has been working well consistently, some limitations still remain. Our experiments use a predefined number of selected samples whereas the traditional active learning may have more flexibility to choose the number of 'valuable' samples to be included. A more thorough exploration of the automatic number and sample selection needs to be done in future work.

As our approach is supervised, it is also interesting to see how it is compared to other unsupervised domain adaptation approaches. We would also like to see the possibility of employing RCA selection without the need of any annotated target database for unsupervised domain adaptation.

Learning between different modalities

There is not much work on multi-modal learning on unpaired images, particularly in medical imaging. Hence, it is difficult for us to compare our proposed approach with existing methods. The limitation of our dual-stream network is that we only evaluate it on single-task in multi-modal learning.

As our future work, we would like to study the power of learning shared representations in multi-modal learning from different datasets with multi-task learning. This study may be beneficial as we rarely have paired images with the same task in different imaging modalities, however, the number of generated medical images are keep rising.

Learning to segment small organs

The limitation of our proposed two-stage weighted network is that it is a non end-to-end approach, hence is more time-consuming. Although the results provide better results on small organ segmentation, one thing that could be improved: the assignment of weight per class in weighted networks could be automatically chosen or learned.

Since our approach is not end-to-end, in the future work, it will be interesting to solve small organ segmentation with an end-to-end approach with automatic network weighting and auto-focus to a particular context. As the existing end-to-end approach [112] is still not satisfactory on small organ accuracy, further study could be directed to improve this.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Evaluating reinforcement learning agents for anatomical landmark detection. *MIDL (Medical Imaging with Deep Learning)*, 2018.

[3] P. Aljabar, R. Wolz, and D. Rueckert. *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*, chapter Manifold Learning for Medical Image Registration, Segmentation, and Classification. IGI Global, 2012.

[4] Ehsan Mohammady Ardehaly and Aron Culotta. Domain adaptation for learning from label proportions using self-training. In *IJCAI*, pages 3670–3676, 2016.

[5] Kolawole O Babalola, Brian Patenaude, Paul Aljabar, Julia Schnabel, David Kennedy, William Crum, Stephen Smith, Tim F Cootes, Mark Jenkinson, and Daniel Rueckert. Comparison and evaluation of segmentation techniques for subcortical structures in brain

mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 409–416. Springer, 2008.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] Wenjia Bai, Wenzhe Shi, Declan P O'Regan, Tong Tong, Haiyan Wang, Shahnaz Jamil-Copley, Nicholas S Peters, and Daniel Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images. *IEEE Transactions on Medical Imaging*, 32(7):1302–1315, 2013.

[8] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M Lee, Nay Aung, Elena Lukaschuk, Mihir M Sanghvi, et al. Human-level CMR image analysis with deep fully convolutional networks. *arXiv preprint arXiv:1710.09289*, 2017.

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CVPR Tutorial*, 2017.

[10] Andrea Baraldi, Lorenzo Bruzzone, and Palma Blonda. Quality assessment of classification and cluster maps without ground truth knowledge. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):857–873, 2005.

[11] Yoshua Bengio et al. Deep learning of representations for unsupervised and transfer learning. *ICML Unsupervised and Transfer Learning*, 27:17–36, 2012.

[12] Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas L Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Advances in neural information processing systems*, pages 177–184, 2004.

[13] Sylvain Bouix, Marcos Martin-Fernandez, Lida Ungar, Motoaki Nakamura, Min-Seong Koo, Robert W McCarley, and Martha E Shenton. On evaluating brain tissue classifiers without a ground truth. *NeuroImage*, 36(4):1207–1224, 2007.

[14] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017.

[15] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient ND image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.

[16] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[17] Robert W Brown, E Mark Haacke, Y-C Norman Cheng, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[18] Lorenzo Bruzzone and Mattia Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

[19] Jinzheng Cai, Le Lu, Yuanpu Xie, Fuyong Xing, and Lin Yang. Improving deep pancreas segmentation in ct and mri images via recurrent neural contextual learning and direct loss function. *arXiv preprint arXiv:1707.04912*, 2017.

[20] Dean Cerrato, Rosie Jones, and Avinash Gupta. Classification of proxy labeled examples for marketing segment generation. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 343–350. ACM, 2011.

[21] Juan J Cerrolaza, Ronald M Summers, and Marius George Linguraru. Soft multi-organ shape models via generalized pca: A general framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 219–228. Springer, 2016.

[22] Sebastien Chabrier, Bruno Emile, Christophe Rosenberger, and Helene Laurent. Unsupervised performance evaluation of image segmentation. *EURASIP Journal on Applied Signal Processing*, 2006:217–217, 2006.

[23] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[24] Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal MR synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814, 2018.

[25] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[26] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *arXiv preprint arXiv:1804.06353*, 2018.

[27] Patrick Ferdinand Christ, Mohamed Ezzeldin A Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin DAnastasi, et al. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 415–423. Springer, 2016.

[28] Patrick Ferdinand Christ, Florian Ettlinger, Georgios Kaissis, Sebastian Schlecht, Freba Ahmaddy, Felix Grün, Alexander Valentinitsch, Seyed-Ahmad Ahmadi, Rickmer Braren, and Bjoern Menze. Survivalnet: Predicting patient survival from diffusion weighted magnetic resonance images using cascaded fully convolutional and 3d convolutional neural networks. In *IEEE 14th International Symposium onBiomedical Imaging (ISBI 2017)*, pages 839–843, 2017.

[29] Paulo Lobato Correia and Fernando Pereira. Stand-alone objective segmentation quality evaluation. *EURASIP Journal on Applied Signal Processing*, 2002(1):389–400, 2002.

[30] Antonio Criminisi, Duncan Robertson, Ender Konukoglu, Jamie Shotton, Sayan Pathak, Steve White, and Khan Siddiqui. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical Image Analysis*, 17(8):1293–1303, 2013.

[31] Antonio Criminisi and Jamie Shotton. *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.

[32] Antonio Criminisi, Jamie Shotton, and Stefano Bucciarelli. Decision forests with long-range spatial context for organ localization in ct volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 69–80, 2009.

[33] Antonio Criminisi, Jamie Shotton, Ender Konukoglu, et al. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

[34] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

[35] Xiang Deng, Lei Zhu, Yiyong Sun, Chenyang Xu, Lan Song, Jiuhong Chen, Reto D Merges, Marie-Pierre Jolly, Michael Suehling, and Xiaodong Xu. On simulating subjective evaluation using combined objective metrics for validation of 3D tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 977–984. Springer, 2007.

[36] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[37] W Thomas Dixon. Simple proton spectroscopic imaging. *Radiology*, 153(1):189–194, 1984.

[38] Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 668–675, 2013.

[39] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 691–697, 2018.

[40] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[41] Wei Fan and Ian Davidson. Reverse testing: an efficient framework to select amongst classifiers under sample selection bias. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 147–156. ACM, 2006.

[42] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sebastien Ourselin, and Tom Vercauteren. Scalable multimodal convolutional networks for brain tumour segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 285–293. Springer, 2017.

[43] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[44] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[45] Kambiz Frounchi, Lionel C Briand, Leo Grady, Yvan Labiche, and Rajesh Subramanyan. Automating image segmentation verification and validation by learning test oracles. *Information and Software Technology*, 53(12):1337–1348, 2011.

[46] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[47] Romane Gauriau, Rémi Cuingnet, David Lesage, and Isabelle Bloch. Multi-organ localization with cascaded global-to-local regression and shape prior. *Medical Image Analysis*, 23(1):70–83, 2015.

[48] Feng Ge, Song Wang, and Tiecheng Liu. New benchmark for image segmentation evaluation. *Journal of Electronic Imaging*, 16(3):33011, 2007.

[49] Ezequiel Geremia, Darko Zikic, Olivier Clatz, BH Menze, Ben Glocker, Ender Konukoglu, Jamie Shotton, OM Thomas, SJ Price, Tilak Das, et al. Classification forests for semantic segmentation of brain lesions in multi-channel MRI. In *Decision Forests for Computer Vision and Medical Image Analysis*, pages 245–260. Springer, 2013.

[50] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. *arXiv preprint arXiv:1702.07841*, 2017.

[51] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2551–2559, 2015.

[52] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[53] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal CT with dense v-networks. *IEEE Transactions on Medical Imaging*, 37(8):1822–1834, 2018.

[54] Ben Glocker, Ender Konukoglu, Ioannis Lavdas, Juan Eugenio Iglesias, Eric O Aboagye, Andrea G Rockall, and Daniel Rueckert. Correction of fat-water swaps in dixon MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 536–543. Springer, 2016.

[55] Michael Goetz, Christian Weber, Franciszek Binczyk, Joanna Polanska, Rafal Tarnawski, Barbara Bobek-Billewicz, Ullrich Koethe, Jens Kleesiek, Bram Stieltjes, and Klaus H Maier-Hein. DALSA: domain adaptation for supervised learning from sparsely annotated mr images. *IEEE Transactions on Medical Imaging*, 35(1):184–196, 2016.

[56] Lutz Goldmann, Tomasz Adamek, Peter Vajda, Mustafa Karaman, Roland Mörzinger, Eric Galmar, Thomas Sikora, Noel E O?Connor, Thien Ha-Minh, Touradj Ebrahimi, et al. Towards fully automatic image segmentation evaluation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 566–577. Springer, 2008.

[57] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016.

[58] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[59] Leo Grady, Vivek Singh, Timo Kohlberger, Christopher Alvino, and Claus Bahlmann. Automatic segmentation of unknown objects, with application to baggage security. In *European Conference on Computer Vision (ECCV)*, pages 430–444. Springer, 2012.

[60] Ricardo Guerrero, Robin Wolz, AW Rao, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative (ADNI), et al. Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO. *NeuroImage*, 94:275–286, 2014.

[61] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[62] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 469–477. Springer, 2016.

[63] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.

[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[65] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3D medical image segmentation: a review. *Medical Image Analysis*, 13(4):543–563, 2009.

[66] Benjamin Hou, Amir Alansary, Steven McDonagh, Alice Davidson, Mary Rutherford, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. Predicting slice-to-volume transformation in presence of arbitrary subject motion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 296–304, 2017.

[67] Yawen Huang, Ling Shao, and Alejandro F Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. *arXiv preprint arXiv:1705.02596*, 2017.

[68] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical Image Analysis*, 24(1):205–219, 2015.

[69] Oscar Jimenez-del Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: Visceral anatomy benchmarks. *IEEE Transactions on Medical Imaging*, 35(11):2459–2475, 2016.

[70] Cheng-Bin Jin, Wonmo Jung, Seongsu Joo, Ensik Park, Ahn Young Saem, In Ho Han,

Jae Il Lee, and Xuenan Cui. Deep CT to MR synthesis using paired and unpaired data. *arXiv preprint arXiv:1805.10790*, 2018.

[71] Thomas Joyce, Agisilaos Chartsias, and Sotirios A Tsaftaris. Robust multi-modal mr image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 347–355. Springer, 2017.

[72] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 597–609. Springer, 2017.

[73] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 587–609, 2017.

[74] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[75] Meina Kan, Shiguang Shan, and Xilin Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3846–3854, 2015.

[76] Timo Kohlberger, Vivek Singh, Chris Alvino, Claus Bahlmann, and Leo Grady. Evaluating segmentation error without ground truth. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–536. Springer, 2012.

[77] Ender Konukoglu, Ben Glocker, Antonio Criminisi, and Kilian M Pohl. WESD-weighted

spectral distance for measuring shape dissimilarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2012.

[78] Ender Konukoglu, Ben Glocker, Dong Hye Ye, Antonio Criminisi, and Kilian M Pohl. Discriminative segmentation-based evaluation through shape dissimilarity. *IEEE Transactions on Medical Imaging*, 31(12):2278–2289, 2012.

[79] Ender Konukoglu, Ben Glocker, Darko Zikic, and Antonio Criminisi. Neighbourhood approximation forests. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 75–82. Springer, 2012.

[80] Ryohei Kuga, Asako Kanezaki, Masaki Samejima, Yusuke Sugano, and Yasuyuki Matsushita. Multi-task learning using multi-modal encoderdecoder networks with shared skip connections. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.

[81] Bart Lamiroy and Tao Sun. Computing precision and recall with missing or uncertain ground truth. In *Graphics Recognition. New Trends and Challenges*, pages 149–162. Springer, 2013.

[82] Bennett Landman, Zhoubing Xu, Juan E. Iglesias, Martin Styner, Thomas R. Langerak, and Arno Klein. Multi-atlas labeling beyond the cranial vault. https://www.synapse.org/Synapse:syn3193805/wiki/217789. doi:10.7303/syn3193805. Accessed June 2017.

[83] Måns Larsson, Yuhang Zhang, and Fredrik Kahl. Robust abdominal organ segmentation using regional convolutional neural networks. In *Scandinavian Conference on Image Analysis*, pages 41–52. Springer, 2017.

[84] Ioannis Lavdas, Ben Glocker, Konstantinos Kamnitsas, Daniel Rueckert, Henrietta Mair, Amandeep Sandhu, Stuart A Taylor, Eric O Aboagye, and Andrea G Rockall. Fully automatic, multiorgan segmentation in normal whole body magnetic resonance imaging, using classification forests, convolutional neural networks, and a multi-atlas approach. *Medical physics*, 44(10):5210–5220, 2017.

[85] Nathan Lay, Neil Birkbeck, Jingdan Zhang, and S Kevin Zhou. Rapid multi-organ segmentation using context integration and discriminative models. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 450–462. Springer, 2013.

[86] Christian Ledig, Wenzhe Shi, Wenjia Bai, and Daniel Rueckert. Patch-based evaluation of image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3065–3072, 2014.

[87] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

[88] Hui Li, Jianfei Cai, Thi Nhat Anh Nguyen, and Jianmin Zheng. A benchmark for semantic image segmentation. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2013.

[89] Xiang Li, Ben Aldridge, Jonathan Rees, and Robert Fisher. Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation. In *Proc. Medical Image Understanding and Analysis Conference*, pages 101–106, 2010.

[90] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[91] Yiming Liu, Jue Wang, Sunghyun Cho, Adam Finkelstein, and Szymon Rusinkiewicz. A no-reference metric for evaluating the quality of motion deblurring. *ACM Trans. Graph.*, 32(6):175–1, 2013.

[92] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[93] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[94] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[95] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[96] Andrzej Mackiewicz and Waldemar Ratajczak. Principal components analysis (PCA). *Computers and Geosciences*, 19:303–342, 1993.

[97] Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötzel, et al. Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. *Computer Vision and Image Understanding*, 164:92–102, 2017.

[98] Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 478–486. Springer, 2016.

[99] Tanmoy Mukherjee, Makoto Yamada, and Timothy M Hospedales. Deep matching autoencoders. *arXiv preprint arXiv:1711.06047*, 2017.

[100] Frank H Netter. *Atlas of Human Anatomy, Professional Edition E-Book: including NetterReference. com Access with Full Downloadable Image Bank*. Elsevier Health Sciences, 2014.

[101] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[102] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 417–425. Springer, 2017.

[103] Dong Nie, Li Wang, Yaozong Gao, and Dinggang Sken. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1342–1345. IEEE, 2016.

[104] Ozan Oktay, Wenjia Bai, Matthew Lee, Ricardo Guerrero, Konstantinos Kamnitsas, Jose Caballero, Antonio de Marvao, Stuart Cook, Declan ORegan, and Daniel Rueckert. Multi-input cardiac image super-resolution using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 246–254. Springer, 2016.

[105] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Ricardo Guerrero, Stuart Cook, Antonio de Marvao, Declan O'Regan, et al. Anatomically constrained neural networks (acnn): Application to cardiac image enhancement and segmentation. *IEEE Transaction of Medical Imaging*, 2017.

[106] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[107] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1717–1724, 2014.

[108] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[109] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[110] Nick Pawlowski, Sofia Ira Ktena, Matthew C.H. Lee, Bernhard Kainz, Daniel Rueckert, Ben Glocker, and Martin Rajchl. DLTK: State of the art reference implementations for deep learning on medical images. In *Medical Imaging meet NIPS Workshop*, 2017.

[111] Steffen E Petersen, Paul M Matthews, Fabian Bamberg, David A Bluemke, Jane M Francis, Matthias G Friedrich, Paul Leeson, Eike Nagel, Sven Plein, Frank E Rademakers, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46, 2013.

[112] Yao Qin, Konstantinos Kamnitsas, Siddarth Ancha, Jay Nanavati, Garrison Cottrell, Antonio Criminisi, and Aditya Nori. Autofocus layer for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 603–611. Springer, 2018.

[113] Prabhu Ramachandran and Gaël Varoquaux. Mayavi: 3d visualization of scientific data. *Computing in Science & Engineering*, 13(2):40–51, 2011.

[114] Andrea Rockall. *Development and evaluation of machine learning methods in whole body MR with diffusion weighted imaging for staging of patients with cancer. Short title: MAchine Learning In whole Body Oncology (MALIBO)*. National Institute of Health Research (UK), 2015.

[115] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[116] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated

pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 556–564. Springer, 2015.

[117] Holger R Roth, Le Lu, Amal Farag, Andrew Sohn, and Ronald M Summers. Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 451–459. Springer, 2016.

[118] Holger R Roth, Hirohisa Oda, Yuichiro Hayashi, Masahiro Oda, Natsuki Shimizu, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. Hierarchical 3d fully convolutional networks for multi-organ segmentation. *arXiv preprint arXiv:1704.06382*, 2017.

[119] Daniel Rueckert and Paul Aljabar. Nonrigid registration of medical images: Theory, methods, and applications [applications corner]. *IEEE Signal Processing Magazine*, 27(4):113–119, 2010.

[120] Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

[121] Gerwin Schmidt, Dietmar Dinter, Maximilian F Reiser, and Stefan O Schoenberg. The uses and limitations of whole-body magnetic resonance imaging. *Deutsches Ärzteblatt International*, 107(22):383, 2010.

[122] Euclid Seeram. *Computed Tomography-E-Book: Physical Principles, Clinical Applications, and Quality Control*. Elsevier Health Sciences, 2015.

[123] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I Sánchez, and Bram van Ginneken. Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[124] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 806–813, 2014.

[125] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.

[126] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[127] Karan Sikka and Thomas M Deserno. Comparison of algorithms for ultrasound image segmentation without ground truth. In *SPIE Medical Imaging*, pages 76271C–76271C. International Society for Optics and Photonics, 2010.

[128] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[129] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 966–973. IEEE, 2010.

[130] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[131] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *ICLR 2017 workshop*, 2017.

[132] Abdel A Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015.

[133] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[134] Tong Tong, Robin Wolz, Zehan Wang, Qinquan Gao, Kazunari Misawa, Michitaka Fujiwara, Kensaku Mori, Joseph V Hajnal, and Daniel Rueckert. Discriminative dictionary learning for abdominal multi-organ segmentation. *Medical Image Analysis*, 23(1):92–104, 2015.

[135] Kuan-Lun Tseng, Yen-Liang Lin, Winston Hsu, and Chung-Yang Huang. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. *CVPR*, pages 3739–3746, 2017.

[136] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757, 2010.

[137] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[138] Sila Ulus, Erdogan Suleyman, Umit Aksoy Ozcan, and Ercan Karaarslan. Whole-body MRI screening in asymptomatic subjects; preliminary experience and long-term follow-up findings. *Polish Journal of Radiology*, 81:407, 2016.

[139] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, 2007.

[140] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*, 36(8):1597–1606, 2017.

[141] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.

[142] Bram Van Ginneken, Tobias Heimann, and Martin Styner. 3d segmentation in the clinic: A grand challenge. *Proceedings of MICCAI Workshop on 3D Segmentation in the Clinic*, pages 7–15, 2007.

[143] Annegreet Van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen De Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Transactions on Medical Imaging*, 34(5):1018–1030, 2015.

[144] Annegreet van Opbroek, Meike W Vernooij, M Arfan Ikram, and Marleen de Bruijne. Weighting training images by maximizing distribution similarity for supervised segmentation across scanners. *Medical Image Analysis*, 24(1):245–254, 2015.

[145] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2001.

[146] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. *CVPR*, pages 5239–5247, 2017.

[147] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[148] Robin Wolz, Paul Aljabar, Joseph V Hajnal, Jyrki Lötjönen, Daniel Rueckert, Alzheimers Disease Neuroimaging Initiative, et al. Nonlinear dimensionality reduction combining mr imaging with non-imaging information. *Medical Image Analysis*, 16(4):819–830, 2012.

[149] Robin Wolz, Chengwen Chu, Kazunari Misawa, Kensaku Mori, and Daniel Rueckert. Multi-organ abdominal ct segmentation using hierarchically weighted subject-specific atlases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 10–17. Springer, 2012.

[150] Xiaojun Wu, Xiaohao Chen, and Jinghui Zhou. C-cnn: Cascaded convolutional neural network for small deformable and low contrast object localization. In *CCF Chinese Conference on Computer Vision*, pages 14–24. Springer, 2017.

[151] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A Bernal, and Jiebo Luo. Deep multimodal representation learning from temporal data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5447–5455.

[152] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2142–2150, 2015.

[153] Varduhi Yeghiazaryan and Irina Voiculescu. An overview of current evaluation methods used in medical image segmentation. Technical report, Tech. Rep. CS-RR-15-08, Department of Computer Science, University of Oxford, Oxford, UK, 2015.

[154] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016.

[155] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *CVPR*, pages 8280–8289, 2018.

[156] Hui Zhang, Sharath Cholleti, Sally A Goldman, and Jason E Fritts. Meta-evaluation of image segmentation using machine learning. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1138–1145. IEEE, 2006.

[157] Hui Zhang, Jason E Fritts, and Sally A Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280, 2008.

[158] Shan Zhao and Qian Zhao. Image retrieval based on mutual information. In *6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)*, pages 1–3. IEEE, 2010.

[159] Yefeng Zheng. Cross-modality medical image detection and segmentation by transfer learning of shapel priors. In *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 424–427. IEEE, 2015.

[160] Yefeng Zheng, Bogdan Georgescu, and Dorin Comaniciu. Marginal space learning for efficient detection of 2d/3d anatomical structures in medical images. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 411–422. Springer, 2009.

[161] Yefeng Zheng, David Liu, Bogdan Georgescu, Daguang Xu, and Dorin Comaniciu. Deep learning based automatic segmentation of pathological kidney in ct: Local versus global image context. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*, pages 241–255. Springer, 2017.

[162] Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2010.

[163] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[164] Xiao-Yun Zhou, Mali Shen, Celia Riga, Guang-Zhong Yang, and Su-Lin Lee. Focal fcn: Towards small object segmentation with limited training data. *arXiv preprint arXiv:1711.01506*, 2017.

[165] Yuyin Zhou, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Deep supervision for pancreatic cyst segmentation in abdominal ct scans. In *International Conference on*

Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 222–230. Springer, 2017.

[166] Yuyin Zhou, Lingxi Xie, Wei Shen, Yan Wang, Elliot K Fishman, and Alan L Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 693–701. Springer, 2017.

[167] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE conference on computer vision and pattern recognition, Hawaii*, pages 7340–7349, 2017.

[168] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[169] Darko Zikic, Ben Glocker, and Antonio Criminisi. Atlas encoding by randomized forests for efficient label propagation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 66–73. Springer, 2013.

[170] Darko Zikic, Ben Glocker, and Antonio Criminisi. Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. *Medical Image Analysis*, 18(8):1262–1273, 2014.

[171] Sonja Zillner, Sascha Seifert, Marius Erdt, Philipp Daumke, and Martin Kramer. Semantic processing of medical data. In *Towards the Internet of Services: The THESEUS Research Program*, pages 343–356. Springer, 2014.

[172] Vasileios Zografos, Alexander Valentinitsch, Markus Rempfler, Federico Tombari, and Bjoern Menze. Hierarchical multi-organ segmentation without registration in 3d abdominal ct images. In *International MICCAI Workshop on Medical Computer Vision*, pages 37–46. Springer, 2015.

[173] Kelly H Zou, Simon K Warfield, Aditya Bharatha, Clare MC Tempany, Michael R Kaus, Steven J Haker, William M Wells, Ferenc A Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index 1: Scientific reports. *Academic Radiology*, 11(2):178–189, 2004.