University of Nottingham
UK | CHINA | MALAYSIA

Semi-supervised Learning for Medical Image Segmentation

Submitted July 2023, in partial fulfillment of
the conditions for the award of the degree **Doctor of Philosophy.**

Ruizhe Li
14342081

Supervised by
Xin Chen
Christian Wagner
Dorothee Auer

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature 李睿哲

Date 26 / 10 / 2023

To my parents and lovely family.

Abstract

Medical image segmentation is a fundamental step in many computer aided clinical applications, such as tumour detection and quantification, organ measurement and feature learning, etc. However, manually delineating the target of interest on medical images (2D and 3D) is highly labour intensive and time-consuming, even for clinical experts. To address this problem, this thesis focuses on exploring and developing solutions of interactive and fully automated methods to achieve efficient and accurate medical image segmentation.

First of all, an interactive semi-automatic segmentation software is developed for the purpose of efficiently annotating any given medical image in 2D and 3D. By converting the segmentation task into a graph optimisation problem using Conditional Random Field, the software allows interactive image segmentation using scribbles. It can also suggest the best image slice to annotate for segmentation refinement in 3D images. Moreover, an "one size for all" parameter setting is experimentally determined using different image modalities, dimensionalities and resolutions, hence no parameter adjustment is required for different unseen medical images. This software can be used for the segmentation of individual medical images in clinical applications or can be used as an annotation tool to generate training examples for machine learning methods. The software can be downloaded from bit.ly/interactive-seg-tool.

The developed interactive image segmentation software is efficient, but annotating a large amount of images (hundreds or thousands) for fully supervised machine learning to achieve automatic segmentation is still time-consuming. Therefore, a semi-supervised image segmentation method is developed to achieve fully automatic segmentation by training on a small number of annotated images. An ensemble learning based method is proposed, which is an encoder-decoder based Deep Convolutional Neural Network (DCNN). It is initially trained using a few annotated training samples. This initially

trained model is then duplicated as sub-models and improved iteratively using random subsets of unannotated data with pseudo masks generated from models trained in the previous iteration. The number of sub-models is gradually decreased to one in the final iteration. To the best of our knowledge, this is the first use of ensemble learning and DCNN to achieve semi-supervised learning. By evaluating it on a public skin lesion segmentation dataset, it outperforms both the fully supervised learning method using only annotated data and the state-of-the-art methods using similar pseudo labelling ideas.

In the context of medical image segmentation, many targets of interest have common geometric shapes across populations (e.g. brain, bone, kidney, liver, etc.). In this case, deformable image registration (alignment) technique can be applied to annotate an unseen image by deforming an annotated template image. Deep learning methods also advanced the field of image registration, but many existing methods can only successfully align images with small deformations. In this thesis, an encoder-decoder DCNN based image registration method is proposed to deal with large deformations. Specifically, a multi-resolution encoder is applied across different image scales for feature extraction. In the decoder, multi-resolution displacement fields are estimated in each scale and then successively combined to produce the final displacement field for transforming the source image to the target image space. The method outperforms many other methods on a local 2D dataset and a public 3D dataset with large deformations. More importantly, the method is further improved by using segmentation masks to guide the image registration to focus on specified local regions, which improves the performance of both segmentation and registration significantly.

Finally, to combine the advantages of both image segmentation and image registration. A unified framework that combines a DCNN based segmentation model and the above developed registration model is developed to achieve semi-supervised learning. Initially, the segmentation model is pre-trained using a

small number of annotated images, and the registration model is pre-trained using unsupervised learning of all training images. Subsequently, soft pseudo masks of unannotated images are generated by the registration model and segmentation model. The soft Dice loss function is applied to iteratively improve both models using these pseudo labelled images. It is shown that the proposed framework allows both models to mutually improve each other. This approach produces excellent segmentation results only using a small number of annotated images for training, which is better than the segmentation results produced by each model separately. More importantly, once finished training, the framework is able to perform both image segmentation and image registration in high quality.

Acknowledgements

I want to acknowledge the incredible support and assistance I received from my supervisors, colleagues, friends, and family throughout this journey. I am profoundly grateful to each and every one of you for making this achievement possible.

First and foremost, I want to express my heartfelt gratitude to my three supervisors, Xin, Christian, and Dorothee. Your unwavering support and help throughout my PhD journey have been invaluable. Special, Xin, very lucky to have you as my supervisor. It is a great pleasure to work with you. Your kindness and patience have made a positive impact on my experience. And your encouragement has been the most important thing that kept me going throughout my doctoral journey. Thank you so much! I'd like to say thank you to Christian for letting me join the LUCID group and taking us on those enjoyable hiking trips. Those memories have been wonderful highlights of my doctoral journey. I also appreciate your patience and support throughout.

I want to express my gratitude to all my colleagues and friends at IMA and LUCID. Working, learning, and chatting with you all has been an absolute pleasure. I feel fortunate to have met such friendly and amazing colleagues like you. I'm also really thankful to all my Chinese friends in Nottingham. Our academic talks, picnics, and friendly chats have made my doctoral life more exciting and enjoyable. Your presence has added so much joy and colour to my time here. Thank you all!

I want to give a special thanks to Xiaowu. Meeting you at the UCL summer school was a fantastic experience. I am so grateful to have made friends with you. I miss our days of discussing academic matters online, even across oceans. I truly value our collaboration during the CMRxMotion Challenge. Thank you for being an amazing friend!

Above all, I want to express my gratitude to my parents and elders. Thank you for your unwavering care and support, which has enabled me to reach this milestone today. Additionally, I would like to thank my wife. Meeting you

during my doctoral journey has been the most fortunate and meaningful thing. Thank you for being by my side!

List of Publications

1. **R. Li** and X. Chen, 'An Efficient Interactive Multi-label Segmentation Tool for 2D and 3D Medical Images Using Fully Connected Conditional Random Field', Computer Methods and Programs in Biomedicine, vol. 213, p. 106534, 2022. [**Chapter 3**]

2. **R. Li**, D. Auer, C. Wagner, and X. Chen, 'A Generic Ensemble Based Deep Convolutional Neural Network for Semi-Supervised Medical Image Segmentation', in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1168–1172. [**Chapter 4**]

3. **R. Li**, M. Bastiani, D. Auer, C. Wagner, and X. Chen, 'Image Augmentation Using a Task Guided Generative Adversarial Network for Age Estimation on Brain MRI', in Medical Image Understanding and Analysis, 2021, pp. 350–360. [**Data augmentation using a Generative Adversarial Network (GAN) for a regression problem.**]

4. **R. Li** and X. Chen, 'Motion-Related Artefact Classification Using Patch-Based Ensemble and Transfer Learning in Cardiac MRI', in Statistical Atlases and Computational Models of the Heart. Regular page in CMRxMotion Challenge held in 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022, pp. 429–438. [**Expanding the concept of ensemble methods to address a classification problem, winning the first place in the CMRxMotion grant challenge.**]

5. M. Jafari, **R. Li**, Y. Xing, D. Auer, S. Francis, J. Garibaldi and X. Chen, 'FU-Net: Multi-class Image Segmentation Using Feedback Weighted U-Net', in Image and Graphics, 2019, pp. 529–537. [**As the second author, the main contribution was to test and modify the feedback weighting loss specifically designed for imbalanced data.**]

6. G. K. Mahani, **R. Li**, N. Evangelou, S. Sotiropolous, P. S. Morgan, A. P. French and X. Chen, 'Bounding Box Based Weakly Supervised Deep Convolutional Neural Network for Medical Image Segmentation Using an Uncertainty Guided and Spatially Constrained Loss', in 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022, pp. 1–5. [**As the second author, the main contributions were the implementation of test time augmentation and the refinement of uncertainty-based loss.**]

The following publications are from a breast cancer diagnosis project in which I participated. My main contributions include MRI image registration, breast tumour segmentation and texture feature extraction.

1. T. M. A. Abdel-Fatah, R. Webb, **R. Li**, X. Chen, E. Giannotti, D. Auer, J. Walker, P. M. Moseley, A. G. Pockley, G. Ball, I. O. Ellis, E. Rakhah, Z. Hodi, Lee, A. A. Chan, and S. Chan, 'Evidence that neoadjuvant anthracycline based combination chemotherapy (NACT) in breast cancer (BC) induces phenotypical changes which guides the optimal adjuvant therapy', Journal of Clinical Oncology, vol. 37, no. 15_suppl, pp. 590–590, 2019.

2. T. M. A. Abdel-Fatah, X. Chen, **R. Li**, E. Giannotti, D. Auer, J. Walker, J. Lim, A. G. Pockley, G. Ball, E. Rakhah, I. Ellis, A. Chan and S. Chan, 'Abstract P3-07-02: Developing a robust multidimensional molecular, pathological and radiological prognostic index (MPRPI) to evaluate the response to neoadjuvant chemotherapy (NACT) and predict clinical outcome of breast cancer (BC)', Cancer Research, vol. 80, no. 4_Supplement, pp. P3-07-02-P3-07–02, 02 2020.

3. T. M. A. Abdel-Fatah, G. Ball, X. Chen, D. Mehaisi, E. Giannotti, D. Auer, J. Vadakekolathu, **R. Li**, G. Pockley and S. Chan, 'Abstract P1-08-19: Utilising artificial intelligence (AI) for analysing multiplex genomic and magnetic resonance imaging (MRI) data to develop multimodality predictive system for personalised neoadjuvant treatment of breast cancer (BC)', Cancer Research, vol. 82, no. 4_Supplement, pp. P1-08-19-P1-08–19, 02 2022.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Medical image segmentation refers to the process of dividing an image (e.g. microscopy, mammogram, magnetic resonance imaging, computerised tomography, etc.) into regions of interest (ROI). The primary objective of segmentation is to identify and extract specific areas that are related to specific clinical tasks (e.g. tumour detection and measurement). This technique offers the advantage of eliminating redundant information from the image, enabling more accurate analysis of the image data by focusing solely on specific areas.

Nonetheless, medical image segmentation can be a time-consuming task, especially when dealing with 3D images. Clinical experts often need to segment each individual 2D slice of a 3D volume to ensure an accurate segmentation result. This process can be laborious and demanding, requiring significant time and effort from the experts.

With the emergence of deep learning, there have been advancements in medical image segmentation. A well-trained deep learning model can automatically and accurately segment new unseen images. However, achieving a "well-trained" model typically necessitates a substantial quantity of high-quality annotated images in the training process. This poses a great challenge, as it requires collecting a large number of manually annotated dataset from clinical experts, which is often impossible due to their limited time and highly demanded expertise. Therefore, the aim of this research work is to develop

efficient and effective medical image segmentation solutions to minimise the reliance on labour intensive image annotations by human experts.

1.1 Background and Motivation

1.1.1 Manual Segmentation is Time Consuming

To segment a medical image, one of the most accurate methods is manual segmentation by a human expert. In the last decade, many image segmentation methods and tools were developed. These include manual segmentation software (e.g ITK-Snap [10], 3D Slicer [11], etc.) and semi-automatic software based on active contour [12], level sets [13], etc. Although the fully-manual segmentation software is able to produce high-quality segmentation results, it is often extremely time-consuming which easily takes hours for a 3D image. Meanwhile, semi-automatic segmentation methods normally require an initial user input, but they do not allow iterative refinement of the segmentation results if inaccurate. Therefore, interactive image segmentation methods become a promising solution that allows iterative improvement of the segmentation results by human inputs.

1.1.2 Challenges in Machine Learning Using Limited Annotated Data

In practical scenarios, the continuous influx of new data necessitates a time-consuming and laborious manual segmention process, even with the assistance of an efficient semi-automatic segmentation method. To alleviate this challenge, deep learning-based approaches offer a promising solution. By training deep learning models on existing annotated data, these models can automatically annotate new data. However, training a deep learning model to achieve satisfactory performance typically requires a fully-supervised learning on a substantial quantity of high-quality annotated data. Unfortunately, this con-

siderably escalates the expenses associated with model training. To address this issue, this thesis focuses on the development of new methods based on semi-supervised learning, which aims to train a deep learning segmentation model with acceptable performance using a limited amount of annotated data in conjunction with a substantial number of unannotated data.

Once some annotated images are produced by using the manual or interactive segmentation tools, there are various ways to train a model by utilising both annotated and unannotated images. Consistency regularisation, generative-based method and pseudo-labelling are commonly used in deep learning based methods[14].

The teacher-student model, a form of consistency regularisation, involves training a teacher model on annotated data and guiding the student model to produce consistent predictions through regularisation techniques [15]. While this approach is effective for classification tasks, it faces challenges in handling complex segmentation and is sensitive to the performance of the teacher model. If the teacher model performs poorly, it can introduce biases and negatively impact the learning process of the entire model.

Generative models can contribute to semi-supervised learning tasks in two ways. Firstly, they can be used to augment data by generating annotated samples, which helps address the issue of unbalanced distribution in annotated datasets [16]. However, unlike classification tasks where images can be generated with corresponding labels, synthesising images with correct masks for segmentation tasks is more challenging. Secondly, a generative model can be trained on the entire dataset unsupervised for the task of image reconstruction, and then transfer the learned model to an image segmentation task. The assumption is that the model learns the underlying data distribution relevant to the segmentation task. The encoding part of the trained generative model can serve as a feature extraction module for image segmentation tasks. By fine-tuning these models using annotated images, their learned features can be transferred effectively to the image segmentation task [17]. However, there are

potential drawbacks of this approach. Image reconstruction tasks may result in information loss, and the limited supervision for segmentation may introduce bias to the annotated data. In summary, generative models offer potential solutions to semi-supervised learning methods, but synthesising accurate masks for segmentation is challenging. Training a generative model for image reconstruction can help capture data distribution, but caution is needed to mitigate information loss and address the bias introduced by the limited supervision.

Another approach suitable for semi-supervised learning is the pseudo-labelling method [18], which can be applied to both consistency regularisation and generative methods. This technique involves training a predictive model on the annotated training set and using it to assign annotations (labels or masks) to the unannotated training set based on heuristics or rules mixed with the annotated data. Although the generated annotations or masks are often noisy and may not accurately reflect the true annotation, this method can utilise the unannotated data by providing additional training information. Pseudo-labelling allows the model to learn more robust representations and decision boundaries, potentially improving predictive performance. It is a flexible approach that can be easily adapted to different domains and tasks. However, the pseudo-annotations generated from the predictive model can be noisy or incorrect, as there is no quality control during the annotation process. Once a pseudo-annotation is assigned, the model cannot correct errors by itself, which can lead to perpetually incorrect results. To mitigate these issues, incorporating regularisation techniques or ensemble methods can help improve the robustness of the learning process and mitigate the effects of incorrect pseudo-annotations.

In addition to the three aforementioned semi-supervised approaches, there is another approach that can be beneficial for medical image segmentation tasks due to the nature of medical images. Many medical datasets exhibit similar features and structures, such as wrist MRI, cardiac MRI, brain MRI, abdominal CT, etc. Moreover, in certain disease diagnoses, multiple medi-

cal images of the same patient may be available from different time points. In such cases, image registration, which aligns two or multiple images with similar structures into a common template space, can be used for the task of image segmentation. One technique that can be utilised for this purpose is the spatial transformer network (STN)-based unsupervised image registration model [9]. This model generates a displacement field for aligning images. By applying the corresponding displacement field, the mask information from one image can be transferred to the unannotated image, effectively annotating the previously unannotated image. Research has shown that incorporating mask-guided training can further improve the quality of segmentation results [19] [20]. Building upon this approach, the combination of image registration and segmentation networks can mutually enhance each other through joint training [21].

However, the use of image registration networks for image segmentation is still an area of limited research. To achieve joint training for various types of data, it is necessary to develop a general image registration network that can handle both small and large deformations. Similar to the pseudo-labelling method, the integrated image registration and segmentation approach is sensitive to the quality of pseudo-masks predicted by both models. Therefore, implementing quality control mechanisms is necessary to ensure the reliability of this approach.

1.2 Aims and Objectives

Based on the challenges discussed above, the aim of this thesis is to propose efficient and effective solutions for medical image segmentation, spanning from data collection to fully automated segmentation models, with a focus on reducing human labour. This research work achieves this aim through the implementation of an interactive image segmentation software and two semi-supervised learning approaches. To successfully accomplish this aim, the

following specific objectives are addressed:

- To ensure the quality of the deep learning model and alleviate the complexity of segmentation, it is necessary to gather data from the clinical domain and seek input from experts for several segmentation of the dataset. In order to meet the quality standards and enhance efficiency, it is imperative to design an interactive segmentation tool for the initial stage. This tool should enable users to swiftly perform automatic image segmentation based on user prompts. Moreover, it should provide the functionality for users to easily make corrections to any incorrectly segmented regions.

- A generic semi-supervised learning framework for various types of medical images. The framework will be capable of learning fully automatic segmentation by leveraging both annotated and unannotated images. The objective is to achieve comparable results to fully supervised segmentation models, which rely solely on annotated images.

- As motivated in the background section of this chapter, based on the characteristics of medical images to achieve a better few-shot learning, another approach explored in this thesis to achieve semi-supervised image segmentation is the combination of image registration and image segmentation. Two sub-objectives need to be achieved:

  - An unsupervised image registration network that can accurately align images at both small and large scales for robust image alignment purpose. The network is designed to enable learning without the need of any masks or annotated data. Furthermore, the network should be capable of enhancing its performance by incorporating guidance from segmentation masks.

  - A joint model that combines image registration and image segmentation networks. The model should be capable of learning from

a limited number of annotated images to obtain annotation information. Moreover, the model is designed to ensure that both the image registration and segmentation networks progress together instead of regressing together. This means that the model is able to avoid each network learning incorrect information and ensure their collaborative advancement.

1.3 Thesis Structure and Key Contributions

This thesis consists of seven chapters, including introduction, literature review, four technique chapters and conclusions.

In this specific chapter, a concise overview of the background information is presented to identify research gaps. These gaps are then utilised to define the aims and objectives of the study. Finally, this chapter wraps up by emphasising the main contributions made by this research work in the next few paragraphs of this section.

Chapter 2 provides the literature review focusing on image segmentation, semi-supervised learning, and image registration. The section dedicated to image registration emphasised the exploration of research that integrates image registration and image segmentation, which is relevant to the present study.

Chapter 3 serves as the first technical chapter in this thesis, which presents an interactive semi-automatic software. This software provides effective annotation of multiple categories for 2D and 3D medical images. Moreover, it includes an automated recommendation function for annotating the next best slice in 3D, thereby enhancing efficiency in the segmentation process. Utilising this tool, clinical experts can easily create annotations and rectify erroneous segmentation results with a few clicks. The annotations collected from these experts will be valuable for the subsequent chapters.

Chapter 4 introduces a new semi-supervised deep convolutional neural network (DCNN) based on an ensemble learning approach. Initially, the network

is trained using a limited number of annotated images. By incorporating a large set of unannotated images, the model's performance is greatly enhanced. This approach enables the construction of a generic automatic segmentation model even with a dataset that only contains a small number of annotated images. By employing a pseudo-labelling approach, the framework is adaptable to various types of data.

Chapter 5 introduces a novel multi-resolution image registration framework designed to accomplish robust deformable image alignment, even for large deformations. This framework also demonstrates its applicability in the image segmentation task by leveraging registration to segment unseen images. By incorporating a mask-guided loss term, the registration accuracy within the masked region shows a significant improvement. This enhancement proves beneficial in establishing the semi-supervised registration-segmentation framework discussed in chapter 6.

Chapter 6 presents the proposed a semi-supervised registration-segmentation framework designed for the automated segmentation of medical images. This framework employs an iterative optimisation process that leverages a pre-trained segmentation model trained using a limited number of annotated images and an unsupervised image registration model using all images. By integrating a segmentation quality assessment block, both the segmentation model and the registration model undergo iterative improvements, leading to enhanced performance in both image registration and segmentation.

Chapter 7 draws conclusions for this thesis, providing a comprehensive summary of the research conducted and the outcomes achieved. It includes a thorough discussion on the limitations of the current work and proposes future plans to address these limitations. Additionally, it highlights additional research conducted during the PhD period, such as medical quality assessment for classification task and the development of a generative model-based data augmentation approach for age estimation on brain MRI datasets (combining image generation and regression techniques).

Chapter 2

Literature Review

2.1 Introduction

This chapter presents a literature review of two key topics that are related to the research of this thesis: image segmentation and semi-supervised learning. Additionally, it introduces the background of medical image registration to support the integration of registration and segmentation framework proposed in chapter 6. As a brief outline, section 2.2 presents a concise review of image segmentation. Section 2.3 provides an overview of semi-supervised deep learning methods. Section 2.4 explores the technology and research on medical image registration that is relevant to this thesis, along with a brief introduction to the joint model of image registration and segmentation. Lastly, section 2.5 summarises the main findings of this chapter, discusses the limitations of existing work, and sets the stage for the research conducted in this thesis.

2.2 Image Segmentation

Image segmentation is the process of partitioning a digital image into multiple image segments based on the semantics of each region [22]. It can simplify the content of images which is beneficial to image analysis, so, it is widely used in various fields (e.g. road environment segmentation for self-driving (Fig. 2.1)

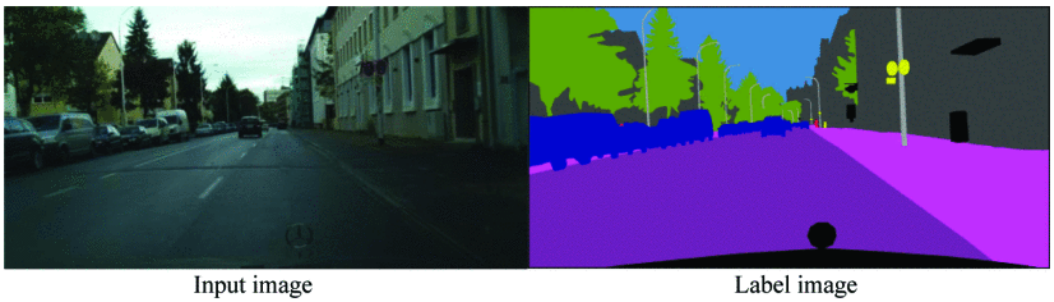

Figure 2.1: *Road semantic segmentation example. From [1] under CC BY 4.0 license.*

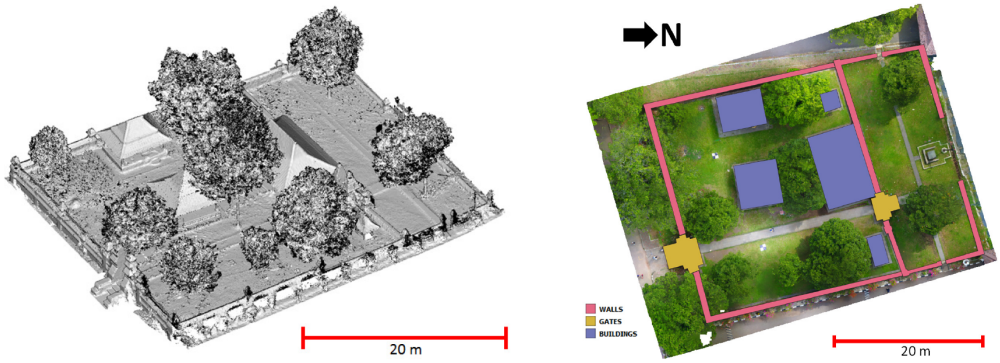

Figure 2.2: *Image segmentation examples for 3D geographic image. Left: original 3D image. Right: segmentation results (red: walls. yellow: gates. blue: buildings). From [2] ©2021 IEEE.*

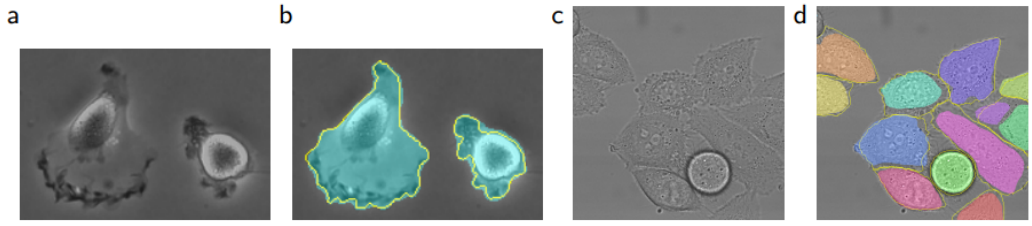

Figure 2.3: *Medical image segmentation examples from [3] ©2015, Springer International Publishing Switzerland. (a) and (c) are the original images, (b) and (d) are the corresponding segmentation results.*

[23] [2], geographic objects segmentation for geographic information system (Fig. 2.2) [1, 24], lesion segmentation for medical imaging (Fig. 2.3) [3]). This thesis mainly focuses on the field of medical imaging.

2.2.1 Classical Methods

Manual segmentation is time-consuming and laborious, so automatic segmentation comes into play. Since the 20th century, many classical automatic segmentation methods have been proposed, such as Ostu's threshold [25], K-

means clustering [26], histogram-based optimal segmentation [27], edge relaxation [28], region growing [29].

More advanced, in some implementations, the image to be segmented is considered as a graph. User's annotations serve as a prior knowledge to determine the likelihood of individual pixel belonging to each of the annotated classes. Together with this prior information, the pixel-wise similarity and label consistency are normally modelled by Markov random filed or conditional random field (CRF). The image segmentation task is then converted into an energy optimisation problem over a graph structure. Although exact inference of such a structure is intractable, a lot of efforts have been made to develop approximation algorithms, including iterated conditional modes [30], belief propagation [31], max-flow/min-cut [32] and filter-based inference [33], in which filter based mean field inference and graph cut are the two most popular solutions.

Boykov et al. [34] proposed the two mostly used graph cut algorithms: $\alpha\beta$-swap and α-expansion. In $\alpha\beta$-swap, for a pair of masks α and β, it exchanges the annotations between an arbitrary set of pixels annotated α and another arbitrary set annotated β. The algorithm generates a mask such that there is no swap move that decreases energy in the predefined graph. The $\alpha\beta$-swap method works well for a binary graph (two-classes) but difficult to be extended to multi-class segmentation. Alternatively, α-expansion is suitable for a multi-class problem. It starts with any mask and runs through all masks iteratively. For each mask α, it computes an optimal α-expansion move and accepts the move if the energy decreases. The algorithm is terminated when there is no expansion move that decreases the energy. Graph cut method has been applied to interactive image segmentation by Rother et al. [35], called "grab cut". In grab cut, users only need to draw a bounding box around the object of interest, the foreground object is then segmented using graph cut. The segmentation result can be further refined using additional scribbles. Grab cut works superbly with minimal user input, but it is limited to binary

class segmentation and the computational speed is slow in 3D. Kohli et al. [36] extended the class of energy functions for which the optimal α-expansion and $\alpha\beta$-swap moves can be computed in polynomial time. However, the inference speed and memory usage is still inefficient comparing to the mean field inference method, especially when there are multiple classes in 3D images.

Many methods of mean field approximations in computer vision have been proposed, such as object class segmentation [37]. The mean field algorithm approximates the exact distribution P using a distribution Q calculated as a product of independent marginal by minimising the KL divergence $D(Q|P)$. Although the approximation of P as a fully factored distribution is likely to lose some information in the distribution, this approximation is computationally efficient. Krähenbühl et. al. [33] developed a filter-based method for performing fast fully connected CRF optimisation, which is the core algorithm used in chapter 3.

These automatic methods are normally application dependent, which can not work robustly unless the object of interest has a homogeneous image intensity and well distinguishable from the other image regions. Many of these methods nowadays are used as a pre-processing step of other more sophisticated methods, such as region of interest extraction for removing redundant information and initial annotation for weakly supervised machine learning methods.

2.2.2 Interactive Image Segmentation

Besides the above classical methods, this section provides a brief overview of popular open-source manual segmentation software. These include both purely manual segmentation techniques and semi-automatic segmentation methods. These tools are more commonly used nowadays than the classical methods, as it ensures the segmentation quality with iterative user interactions.

Many commercial and open-source software offer manual delineations for

medical images, including line tracing, polynomial curve fitting, area painting, etc. ITK-Snap [10] is an open-source and widely used tool that is mainly dedicated to medical image segmentation. It offers polygon and paintbrush tools for flexible editing of both 2D and 3D images. Another widely acknowledged open-source tool is 3D Slicer [11]. It provides manual tools such as area painting, level tracing and scissors, which are normally used as post-processing to refine segmentation results using threshold or region growing. These manual tools offer good quality control but require tremendous time and effort from the user.

Semi-automatic methods take the advantage of automatic segmentation and allow users to intervene with the segmentation process. One type of user interaction is initialisation, such as drawing seeds or bounding boxes inside or around the target object. Then the seeds or initial contour evolve to the desired object's boundary by region growing [38] or minimising an energy function (e.g. active contour [12], level sets [13], etc.). These methods do not offer post-segmentation user interactions to further refine the results and the parameter settings are highly application dependent. Another type of user interaction is to iteratively improve the segmentation results by adding scribbles to different classes (e.g grow cut [39], graph cut [34], etc.). At each iteration, the method propagates these mask to the whole image by optimising an energy function. This is more or less guaranteed to achieve a satisfactory result with reduced workload compared to a manual process, which is desirable for medical image segmentation. These tools are also utilised to provide annotations of a dataset to train deep learning based methods.

2.2.3 Deep Learning based Methods

More recently, many image segmentation methods have been proposed based on deep learning techniques. This section introduces several state-of-the-art deep learning based image segmentation methods.

Region-CNN Based Models

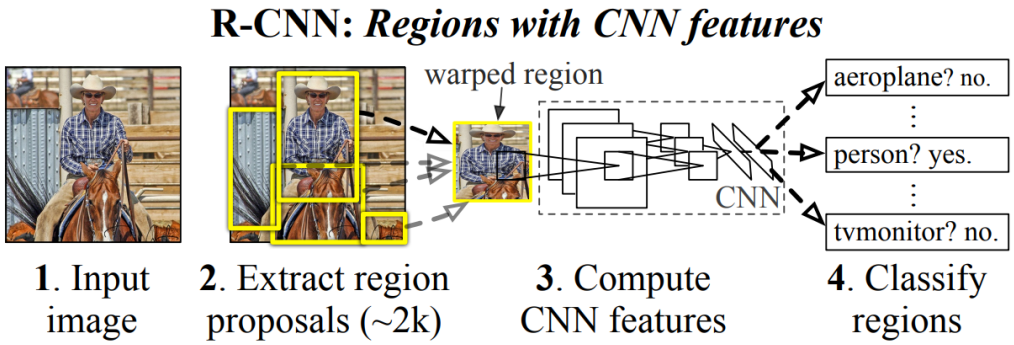

Figure 2.4: *R-CNN overview: Input a image, locate 2000 object candidate bounding-boxes, and then use CNN to extract the feature from each candidate bounding-box, then use classification algorithm to classify and recognise the objects in each candidate bounding-box. From [4] ©2014 IEEE.*

In the end of 2013, Girshick et al. [4] proposed one of the first deep learning based image segmentation network, Regions with Convolutional Neural Network Features (R-CNN). For a given image, the R-CNN segment it through 4 steps (shown in Fig. 2.4): 1) select candidate bounding-boxes using selective search algorithm. 2) extract features by using trained CNN models. 3) classify the object in each candidate bounding-boxes. 4) find the tighter bounding boxes by shrinking the bounding-box to the edge of target using regression model. The method was evaluated on the PASCAL VOC challenge dataset [40] and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [41] [42], outperform other object detection and image segmentation methods at the time. However, The selective search algorithm cannot be GPU accelerated, which reduces the model's speed. Extracting the features using CNN for 2000 candidate bounding-boxes is also not efficient. Additionally, due to the limitation of memory, the model needs to write the image of each candidate bounding-box to the hard disk, which slows down the inference speed. Finally, the whole model is not end-to-end (CNNs extract image features, classification models predict categories, and regression models find tight boundary boxes). Each part is trained separately, which is troublesome to organise the whole structure.

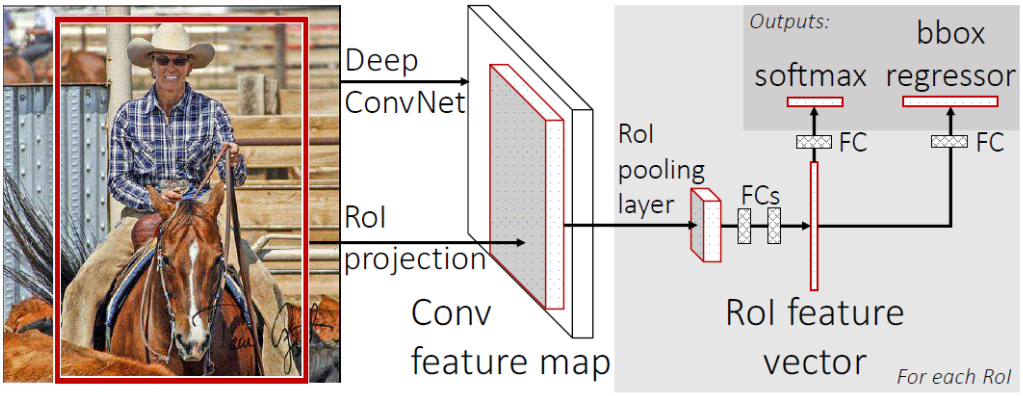

Figure 2.5: *Fast R-CNN overview: An image and multiple regions of interest are input into the full convolutional network. Each RoI is pooled into a fixed-size feature map, which is then mapped to feature vectors via the full connection layer. Two outputs for classification and regression with multi-task loss function achieved end-to-end training. From [5] ©2015 IEEE.*

In the following years, the extensions of R-CNN (Fast R-CNN [5], Faster R-CNN [43] and Mask R-CNN [6]) fill the deficiency of R-CNN and achieve remarkable results in the field of object detection and image segmentation. As shown in Fig. 2.5, instead of send 2000 candidate bounding-boxes to CNN model in R-CNN, the CNN in Fast R-CNN extracts features from an image with multiple regions of interest (RoIs). This greatly reduces the training parameters and significantly improves the training speed. In addition, Fast R-CNN also combines the two loss functions, object classification and the boundary box regression into one, so that they share parameters and train together. It further reduces the number of training parameters, and realises end-to-end training of object detection and segmentation. Moreover, the Faster R-CNN introduces the Regions Proposals Networks (RPN) to select candidate bounding-boxes. A CNN extracts feature maps from input image, the Regions Proposals Networks (RPN) select candidate boxes automatically, then the classification and regression layers local and segment the object. So far, R-CNN achieved single-model full-function end-to-end training.

In 2017, He et al. [6] proposed the Mask R-CNN. This method utilises the previous R-CNN algorithm, combines Faster R-CNN with FCN (described in next sub-section), and obtains excellent results on instance segmentation tasks.

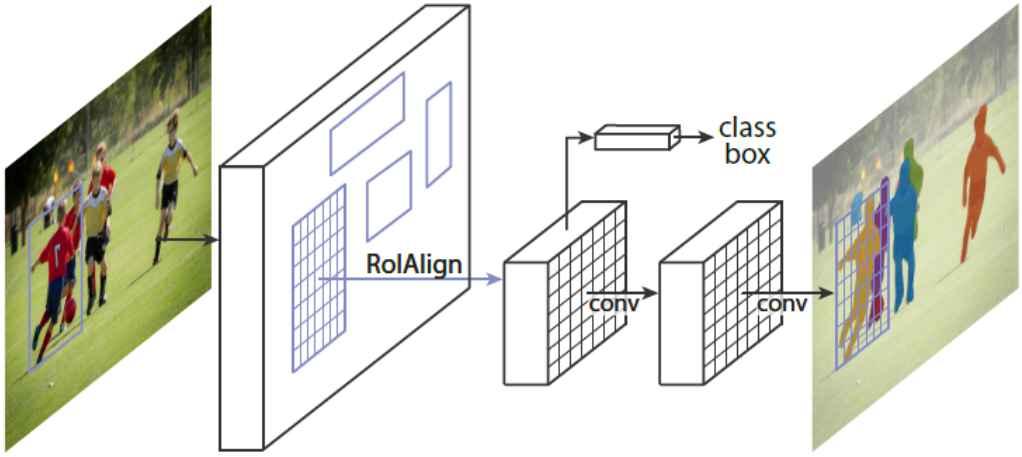

Figure 2.6: *The framework of Mask R-CNN. From [6] ©2017 IEEE.*

As shown in Fig. 2.6, Mask R-CNN has three outputs. The first one is a class label for each object. The second one is the offset for each bounding-box. The third one is a pixel-wise mask for each object from FCN. Thus, Mask R-CNN can predict pixel-wise segmentation results for each instance in the image. The method outperformed all previous methods on different Common Object in Context (COCO) challenge.

R-CNN based methods have shown excellent performance in the field of object detection and achieved good results on instance segmentation tasks, however, there is still room for improvement in the segmentation of object details (semantic segmentation).

Fully Convolutional Networks

In 2015, Long et al. [7] proposed Fully Convolutional Network (FCN) which is one of the first pixel-wise image semantic segmentation method. As shown in Fig.2.7, different from classical CNN, FCN uses the deconvolutional layer instead of the fully connected layer + softmax layer after the feature extraction layers (convolutional layer). The deconvolutional layer up-samples the feature map of the last convolutional layer to restore it to the same size as the input image. Finally, pixel by pixel classification is carried out on the up-sampled feature map, which achieves dense predictions for each pixel. In short, the

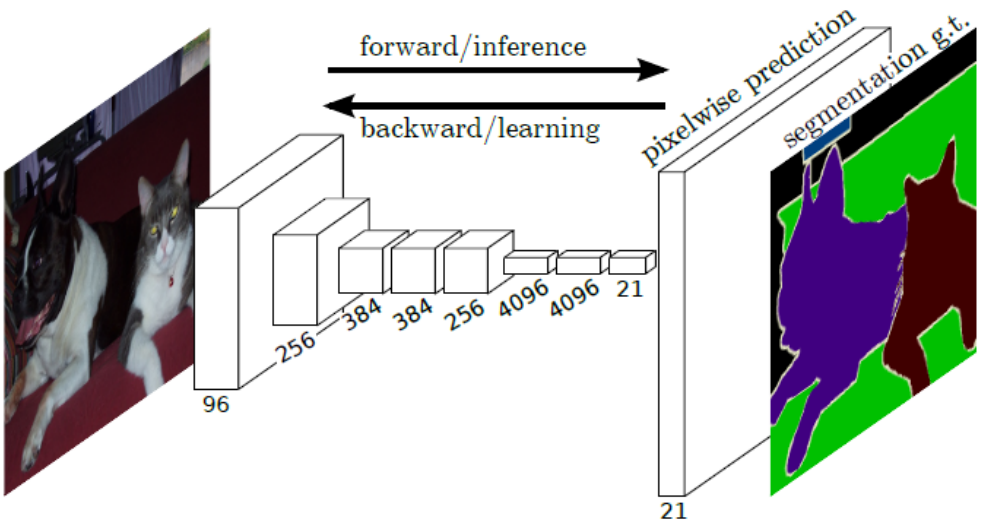

Figure 2.7: *Framework of the Fully Convolutional Network. It can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation. [7] ©2015 IEEE.*

difference between FCN and CNN is that the last fully connected layer of CNN is replaced by the convolutional layer, and the output is a segmented image.

However, if the encoder down-samples the image multiple times, direct amplification of the final feature map back will lead to inaccurate segmentation results. To overcome this problem, the paper proposes a slightly higher precision of the mixed amplification structure. It combines multi-resolution features from coarse to fine. For example, as shown in Fig. 2.8, the lowest precision output FCN-32s is directly obtained from the final convolutional layer conv7. A more precise output FCN-16s is obtained by combining the feature map from pooling layer pool4 and the upsampled convolutional layer conv7. As with the previous method, the output of FCN-8s combines the output of pool3, pool4 and conv7 together, achieves improved segmentation precision. The methods achieved state-of-the-art performance on multiple datasets, including PASCAL VOC, NYUDv2 [44], and SIFT Flow [45].

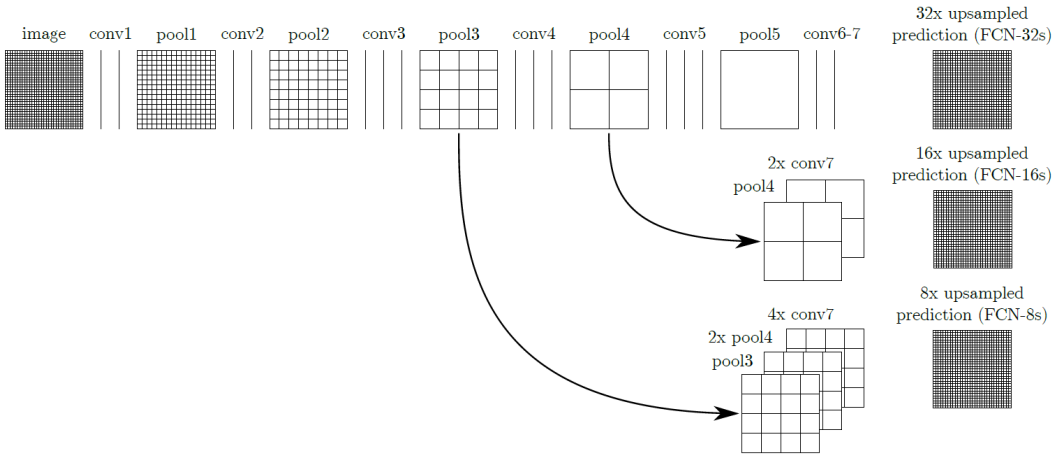

Figure 2.8: *Coarse to fine combination of feature maps for FCN. From [7] ©2017 IEEE.*

Encoder-decoder based Models

To further solve the problem of low precision of FCN, the encoder-decoder based methods were proposed. In 2015, Noh et al. [8] published one of the first encoder-decoder based semantic segmentation network known as DeconvNet. Different from the FCN, an decoder is added to the end of encoder (shown in Fig. 2.9). Inspired by FCN, the encoder using fully convolutional layers adopted from VGG-16 [46]. The decoder is a multi-layer deconvolutional network, which maps the feature vector from encoder to a accurate segmentation map. In this way, the network is able to generate pixel-wise segmentation results for a given image. In the same year, Badrinarayana et al. [47] proposed SegNet which is similar to DeconvNet, but it adds a batch normalisation layer after each convolutional layer and removes the fully connected layer between the encoder and the decoder. Both methods achieved remarkable results on semantic segmentation tasks.

Since then, more and more encoder-decoder structures have been proposed for segmentation, e.g. RefineNet [48], U-Net [3], GCN [49], etc. With the development of deep learning techniques, other structures such as attention mechanisms have been added to encoder-decoder structures to further improve the accuracy of segmentation, such as Chen's scale-aware network [50] and Fu's dual attention network [51]. In recent years, one of the most cutting-

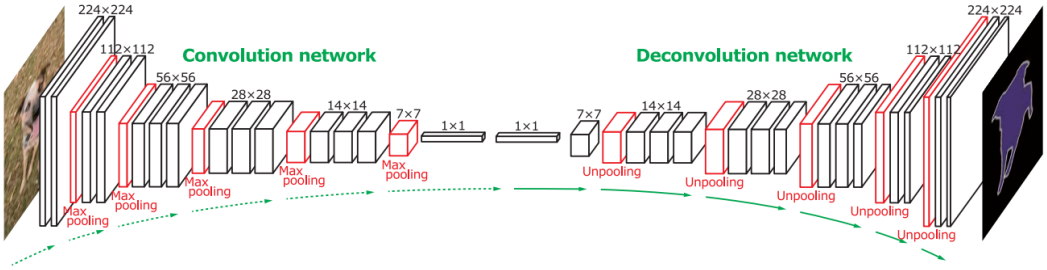

Figure 2.9: *Deconvolutional semantic segmentation. One of the first encoder-decoder based semantic segmentation network. The Encoder is adopted from VGG-16, and the decoder consists with convolutional layer and deconvolutional for up-sampling. Two fully connected layer is used to connect encoder and decoder. From [8] ©2017 IEEE.*

edge technique of attention mechanism, transformer, has been extended from Natural Language Processing (NLP) area to Computer Vision (CV) area, and is widely used in image segmentation tasks. For example, inspired by Vision Transformer (Vit) [52], Strudel et al. [53] proposed the Segmenter, which uses a pure transformer structure to encode and decode the image. The method was evaluated on the challenging ADE20K [54] dataset, and outperformed all previous works.

2.2.4 Medical Image Segmentation

In medical imaging field, image segmentation is a basic and crucial step for many biomedical image analysis tasks (e.g. tumour quantification, cell segmentation, organ analysis, etc). Early approaches for medical image segmentation typically relied on techniques such as edge detection, region growing and traditional machine learning techniques. These methods have achieved good results to some extent, but compared to other natural images, medical images tend to be noisy, blur and low contrast. Hence, medical image segmentation remains one of the most challenging topics in computer vision area. In addition, the two most commonly used medical images, CT (Computed Tomography) and MRI (Magnetic Resonance Imaging), are 3D data, which also make the segmentation more challenging. With the rapid development of deep learning techniques, convolutional neural networks (CNNs) have been successfully im-

plemented based on hierarchical feature representation of the images. CNNs for feature learning provide excellent segmentation results for medical images due to their insensitivity to image noise, quality, contrast, etc.

These deep learning-based methods are widely used in various areas of medical image segmentation, including 2D: cell segmentation [55] [56], skin lesion segmentation [57] [58], retinal vessel segmentation [59] [60], etc., and 3D: cardiac segmentation [61] [62], liver segmentation [63] [64], brain tissue segmentation [65] [66], tumour segmentation [67] [68] [69], etc.

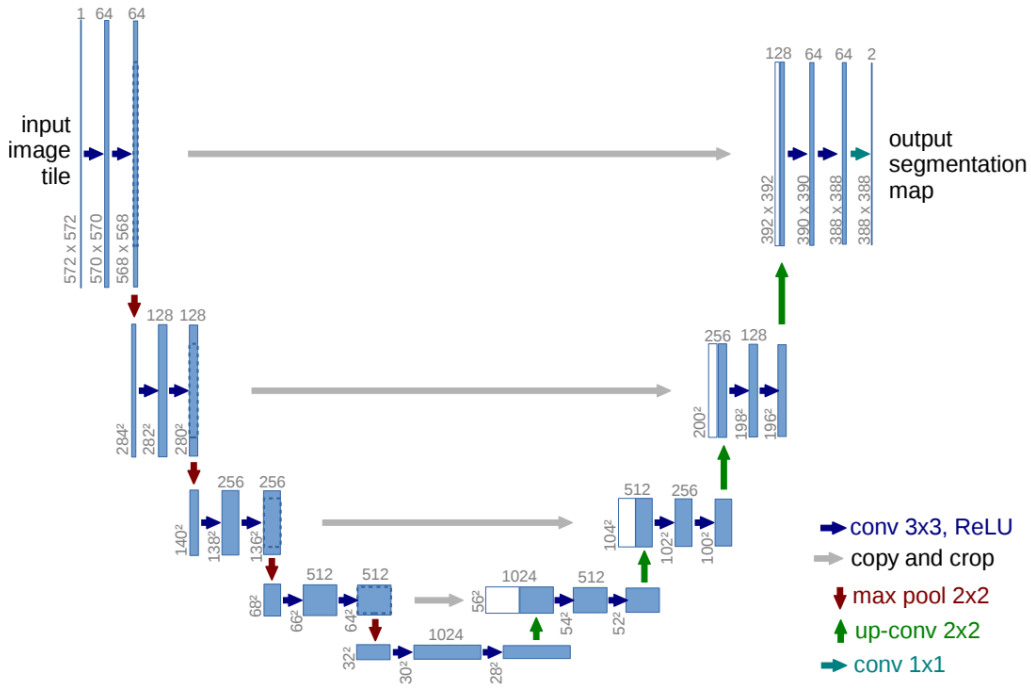

Figure 2.10: *U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. From [3] ©2015, Springer International Publishing Switzerland.*

It is especially worth noting that Ronneberger et al. [3] proposed one of the most popular medical image segmentation network in 2017, known as U-net. As shown in Fig. 2.10, U-net was named based on the shape of the model, which resembles a capital U. Similar to DeconvNet, it has an encoder-decoder structure. The encoder part down-samples the image with convolutional layers to extract features, similar to FCN. The decoder part uses deconvolutional

layers to up-sample the learned feature maps to generate the final pixel-wise segmentation results. Different from the previous encoder-decoder based methods, a skip-connection operation from encoder to decoder is added at every resolution to combine low- and high-resolution feature maps. With this strategy, U-net won the first place in ISBI cell tracking challenge 2015. U-net is also widely used in other medical image segmentation tasks. It has now become the benchmark for most medical image segmentation tasks and has inspired many recent improvements such as 3D U-net [70], V-Net [71], H-DenseUNet [72], TransUNet [73].

Later, Isensee et al. proposed nnU-Net, also known as "no-new-UNet", which incorporates a self-configuring framework for the original U-Nets [74]. nnU-Net is capable of automatically optimising hyper parameters and applying data augmentation strategies without the need for manual intervention. The author demonstrated that nnU-Net achieved top rankings in various public medical image segmentation challenges, and other researchers have also consistently ranked highly using this framework in well-known medical segmentation challenges. This dominance showcases the power of U-Net as a fundamental network for medical image segmentation tasks. Therefore, in this thesis, all deep learning-based medical image segmentation tasks will utilise U-Net as the foundational network.

2.3 Semi-supervised Deep Learning

Nowadays, many encoder-decoder based deep convolutional neural networks (DCNNs) such as U-Net [3] have achieved state-of-the-art performance for image segmentation using fully-supervised learning. However, data annotation is extremely time-consuming especially for medical imaging where highly skilled expertise is required. Several methods have been proposed to address this challenge. Data augmentation is commonly used as an effective solution. A few studies show that geometric transformations and intensity shifts to increase the

number of annotated data can achieve better performance than only using the original annotated data [75]. In this section, another approach is introduced, deep learning based semi-supervised learning, which uses both annotated and unannotated data for deep learning model training. By extracting annotation prediction related information from unannotated data, boost the performance of predictive models.

Depending on the loss function and model design, semi-supervised learning can be classified into various types, including consistency regularisation methods, generative methods, pseudo-labelling methods and graph-based methods [14] [76]. Based on the relevance of the study, this section will only give a brief introduction about generative methods, consistency regularisation methods and pseudo-labelling methods.

2.3.1 Generative Model based Methods

As mentioned above, to ensure effective semi-supervised learning, the model needs to be able to learn information about the annotation predictions from the unannotated data. For a generative model, the key task is to learn and model the real distribution of the training dataset and then generate new data from this distribution.

In this case, one of the most popular generative models, Generative Adversarial Network (GAN) [77], is widely used to generate data that matches the real data distribution. A typical GAN is utilised for generating high-quality images from a random latent vector. It has two parts, a generator G and a discriminator D. It is trained by optimising the following objective function:

$$\min_{G} \max_{D} \mathbb{E}_z \left[D \left(G \left(z \right) \right) \right] - \mathbb{E}_x \left[D \left(x \right) \right] \tag{2.1}$$

where $G : z \rightarrow x$ is the generator that maps an input noise z to its target generated image x. D indicates the discriminator that classifies if an image is real or fake. The generator intends to fool the discriminator by producing

realistic images, and the discriminator aims to identify the fake ones from the real images.

To achieve semi-supervised learning, a simple way is to combine the limited annotated data with synthetic data together to create a combined dataset to train a fully-supervised model. For example, Maayan et al. [16] proposed a data augmentation method which enlarges the size and diversity of the training dataset by adding synthesised images using GAN. It improved the model performance significantly on a liver lesion classification task. Similarly, Qin et al. [78] introduced a GAN-based data augmentation method with style-based GAN architecture. By involving the progressive GAN [79] and the style control technology from styleGAN [80], it achieved a high resolution and rich diversity image generation on a small, complex and class-imbalanced public skin lesion dataset, ISIC 2018 [81] [82]. Then a classification model refines the pre-trained ResNet50[83] model on both real and synthesised data. The approach successfully fills in the imbalances of the original data and delivers remarkable classification results for skin lesion diagnostic tasks.

Generative model based data augmentation has emerged as a promising approach to support semi-supervised learning. However, the effectiveness of this method is limited by the quality of the generated images. As a result, an alternative way to leverage the generative model in semi-supervised learning is to reuse the discriminator for classification tasks. This approach is based on an assumption that the generative model learns the transferable data distribution relevant to the image down-sampling task. In a notable study conducted in 2015, the categorical generative adversarial networks (CatGAN) [84] was proposed by Springenberg, where they integrated the discriminator with a classification function. They then used a classification loss to make the generator generating samples uniformly across all categories such that the discriminator has highly deterministic categories, and make the discriminator classifying the input samples evenly and accurately. Similarly, the semi-supervised learning GAN (SGAN) [17] and the improved GAN [85] were also proposed to address

the semi-supervised learning problem. They modified the output of the discriminator, transforming it from a binary classification of real or fake to a multi-class classification of $[0, 1, 2, ..., K, \text{fake}]$, where K represents the class labels. The researchers trained this new model using both annotated and unannotated data, employing two distinct loss functions. The supervised loss is based on the annotated images and aims to minimise the error in predicting the class labels. On the other hand, the unsupervised loss utilises the unannotated images to distinguish between real and fake images.

However both the above two approaches are for semi-supervised classification tasks, which are difficult to be adopted for image segmentation. The synthetic method has a high risk to generate unmatched image and mask, and add segmentation term is much difficult than adding a classification term to the discriminator. Only a few studies have been conducted on semi-supervised learning tasks using GANs. One such example is a study by Lahiri et al. (2018) [86], where the discriminator in GAN was utilised for both image segmentation and distinguishing between real and generated fake images. By incorporating annotated images, the segmentation accuracy was improved, while the utilisation of unannotated images enhanced the discrimination power. However, it is important to note that this particular method primarily focuses on extracting global information from the images to enhance the discrimination capabilities, rather than emphasising the extraction of segmentation-specific information.

2.3.2 Consistency Regularisation-based Methods

Besides the generative model based methods, another well-known method for semi-supervised learning is to use consistency regularisation. It is based on the assumption that when a very small realistic perturbation is added to a data, the class label of that data should not change. For instance, if one creature is labelled as a “cat” and another creature has the same appearance but the eye colour is green instead of red, it is reasonable to label this creature to “cat” as

well rather than being classified as a completely different category, such as a "dog". In general, consistency regularisation aims to regularise the model by enforcing consistency between different representations or predictions derived from the same input but with different perturbations.

In semi-supervised learning, consistency regularisation is a loss function that focuses on the unlabelled images. By adding small perturbations to the unannotated data, the model is expected to produce consistent outputs. In 2015, Rasmus et al. proposed the Ladder network for semi-supervised image classification by introducing a consistency regularisation loss to the unannotated images [87]. They extended the classification network by adding a noisy encoder and a denoising decoder. In details, given an input image, denoted as x, the Ladder Network produces two outputs: a clean prediction y and a noised output y'. The noised output is generated by injecting Gaussian noise into each layer of the encoder. The denoising decoder then takes the noisy representations from each layer of the encoder as input and reconstructs the original input x. To achieve consistency regularisation for the unlabelled images, the Ladder Network minimises the difference between the original input and the reconstructed input at each layer. This encourages the model to produce consistent predictions and learn robust representations. In combination with a supervised loss computed on the annotated images, this method achieves remarkable results surpassing those obtained when using annotated images alone. In subsequent works, Laine and Aila [88] simplified the Ladder Network architecture by replacing the denoising decoder with a generative model. This modification aimed to streamline the model and reduce its complexity. They then introduced an alternative version, Temporal Ensembling, that incorporated ensemble-based temporal consistency into the model's predictions, utilising temporal information to enhance the accuracy of the predictions. It achieved the state-of-the-art results on various semi-suerpvised learning benchmarks.

Inspired by the Pi-Model and Temporal Ensembling, the Mean Teacher

method was proposed by Tarvainen [15]. It adopts a teacher-student structure, where the teacher model learns from annotated images and instructs the student model by employing consistency regularisation. During training, when provided with the same input, the goal of the student model is to produce the same predictions as the teacher model. The divergence between the two models is optimised through a consistency loss function. Evaluation of this approach on a public classification dataset demonstrates its superior performance compared to Temporal Ensembling. Substantially, the Dual Students model [89] utilises two student models with different initialisation of weights to mitigate bias and enhance prediction stability in compare to the Mean Teacher approach.

The teacher-student model is also able to be adapted to the image segmentation tasks. Cui et al. proposed a semi-supervised teacher-student model for brain lesion segmentation [90]. Similar to the Mean Teacher, the teacher model is trained on labelled dataset, and the student model learns from both the real data and the predicted results of the teacher model. Hang et al. also proposed a similar work but added a local attention to the target region for left atrium segmentation [91]. Zheng et al. added a random Gaussian noise to the student model when updating the teacher model to improve the robustness of the network [92]. Luo et al. conducted an investigation where they employed the Transformer architecture instead of CNN in the teacher-student model for achieving remarkable results on a public benchmark [93]. However, in most implementations of the teacher-student model, consistency regularisation is employed by minimising the discrepancy between the predictions of the teacher model and the student model. This process can be seen as a form of self-learning, where pseudo-masks are utilised. Moreover, the consistency regularisation method used for both image classification and image segmentation tasks is highly sensitive to the annotated dataset. If there is a bias in the annotated data, it can negatively impact the overall performance of the model during the learning process.

2.3.3 Pseudo-labelling Methods

Another effective and widely applicable method for semi-supervised learning is pseudo-labelling [18]. Essentially, this method involves using a predictive model to generate pseudo annotations for unannotated data and then train a model using both the annotated images and the images with pseudo annotations. In this way, the model can learn information from the whole dataset. However, a major drawback of the pseudo-labelling method is that the model cannot correct its own errors. If the model overly confidence in its predictions without acknowledging the potential of inaccuracies, it may result in incorrect outcomes, and the error is propagated during the training process.

To address this issue, an older training strategy called Co-Training was proposed [94]. It requires a dataset where each data has two different views. Two separate models ($M1$, $M2$) are trained on the different views. The data is iteratively added to the other subset based on the model's confidence on its predictions. Specifically, in each iteration, if one of the models (e.g. $M1$) has a high level of confidence in the predicted result for a sample x, $M1$ generates a pseudo-label for that sample and is then included in the training subset of the other model $M2$. As an improvement, democratic co-learning was introduced to address the challenge of collecting different views for every dataset [95]. Instead of relying on different view data, it employs different learning algorithms and avoids bias through majority voting.

For the image segmentation task, how to evaluate the confidence of the predicted results and select the valuable pseudo masks are challenging. Instead, some methods use trusted models to generate pseudo masks. For example, Sun et al. [96] used a teacher model to generate pseudo masks for the task of liver segmentation. Filipiak et al. also used a teacher model to generate the pseudo mask but use bounding boxes and mask scoring to filter out noisy pseudo labels [97]. Feng et al. [98] [99] pointed out that it is difficult for models to reveal their own errors. Instead, exploiting inter-model differences between different

models is the key to locating pseudo-labelling errors. They then proposed the dynamic mutual training (DMT) which trained two different models mutually by dynamically re-weighting the loss function. The methods achieved state-of-the-art performance in both classification tasks and segmentation tasks. Bai et al. [100] developed a self-learning technique which can correct pseudo masks through a post-processing approach. A fully supervised model is firstly trained on annotated data, then pseudo masks are generated for unannotated data using this model and refined by a fully-connected conditional random field (CRF). Subsequently, both annotated data and the data with pseudo masks are used to refine the initial model. This process is repeated until convergence.

Compared to other previously mentioned methods, the pseudo-labelling technique offers a more straightforward approach for semi-supervised image segmentation learning. By generating pseudo masks for the unannotated data and incorporating them into the training process, this method enables the model to leverage a larger dataset and improve performance in a semi-supervised setting. However, careful handling of the pseudo masks is essential to ensure reliable results and mitigate the risk of incorporating erroneous information into the learning process.

2.4 Medical Image Registration

Image registration is a technique that geometrically transforms one image (source) to another (target) image's space, so that the transformed source image is similar and comparable to the target image. The estimated geometric transformation and the warped source image can then be utilised for further analysis. It is a commonly used method in a variety of fields, including medical imaging, remote sensing, computer vision, and robotics. It particularly plays a very important role and has been widely used in the medical field. For instance, aligning medical images (e.g. MRI, CT) that captured from different time points or different subjects for quantitative analysis in disease diagnosis

and prognosis.

Image registration methods can be broadly categorised into two types: rigid registration and non-rigid registration. Rigid registration involves a global match between two images while preserving the original shapes of objectives within the images. This type of registration typically includes operations such as shifting, rotation, and scaling transformations. On the other hand, non-rigid registration allows for local deformations, enabling more complex transformations. It aligns images by utilising a computed deformation field derived from mathematical models or algorithms.

In medical imaging area, both methods motioned above are widely used. Specifically, with the property of rigid transformation preserving, rigid registration are commonly used in alignment of different images of the same patient, such as aligning images from different modalities (e.g. CT and MRI) and aligning images at different time point for the same lesion or organ (e.g. different cycles of cancer).

2.4.1 Classical Image Registration Methods

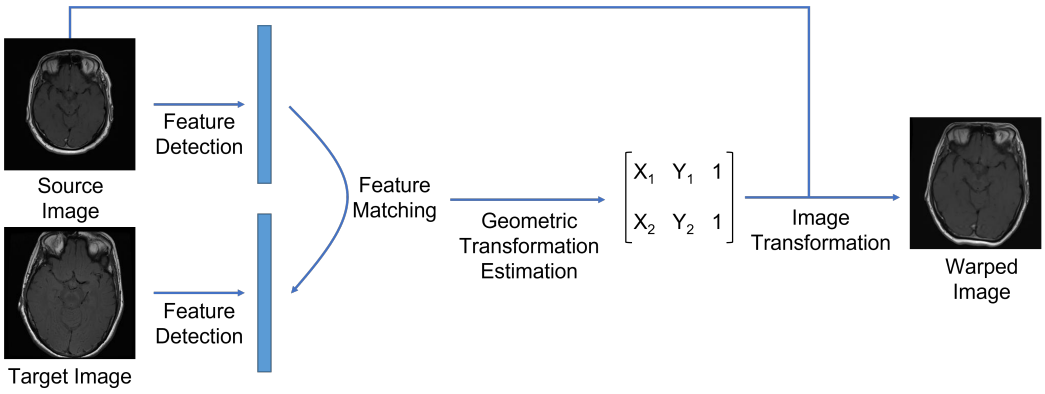

Figure 2.11: *Diagram of image registration.*

Prior to the deep learning era, there are many non-learning-based approaches proposed by researchers to align two or more images. As illustrated in Fig. 2.11, the traditional image registration methods normally include four steps [101]: feature detection, feature matching, geometric transformation estima-

tion and image transformation (warping). There are many popular image registration methods developed based on this pipeline. For example, intensity-based methods using sum of squared difference (SSD) [101], normalised cross-correlation (NCC) [102] and mutual information [103] as the similarity measurements, feature-based method based on scale invariant feature transform (SIFT) [104], point-based method based on iterative closest point (ICP) algorithm [105], etc.

However, these non-learning-based methods normally require careful parameter tuning in each application to achieve satisfactory results, and it is time-consuming to register large images such as 3D medical images. It takes several minutes or more for the alignment of one pair of images.

2.4.2 Deep Learning-based Image Registration Methods

Recently, deep learning methods have achieved remarkable performance in supervised learning. It is able to learn the relation between the known input x and the target result y that aims to predict y for a given x. In image registration task, the paired source image and target image can be considered as two inputs x_1 and x_2 and the geometric transformation is the target y. Therefore, with a set of given paired x_1, x_2 and y, a supervised deep learning network can be trained. Salehi et al. [106] proposed a deep rigid registration method using this idea. They apply the rigid transformations, including random rotations and translations, to the source image x_1 to get the simulated target image x_2, the transformation metric here is the y in the model training. With joint losses, mean squared difference and geodesic distance, the method achieved a remarkable performance in a 3D brain MRI dataset. Instead of rigid transformation, Sun et al. proposed the DVNet [107] which collects a large set of artificially generated displacement vectors (DVs) by expert. Then a fully convolutional neural network (CNN) was trained to estimate the DVs by giving

paired source and target images. The method is able to produce robust results on single-modal liver data, and also works well on simulated CT-US data. However, it did not work on the real CT-US data due to the large appearance differences between the real and simulated images.

Supervised deep learning image registration methods have greatly changed and enhanced the accuracy and effectiveness of image registration compared to the non-learning based techniques. These methods need a training process that involves using a group of example images and a transformation matrices, which is done offline. When registering new images, the computational speed is much faster (usually in seconds) compared to the traditional methods, while still achieving high accuracy. However, it's difficult to obtain the exact transformation matrix or displacement field from real data, which poses a new challenge in training the model without reliable ground truths.

In 2015, Jaderberg et al introduced the spatial transformer network (STN) [9], which enabled data to be manipulated spatially within the network. The STN can be easily added to an existing CNN module without altering the model training process. By using the STN, the model can learn how to deform or transform images by studying paired source and target images, without needing an exact transformation measurement as a reference. In detail, a displacement field is utilised by the STN to warp the source image. The warped source image is then compared to the target image in order to calculate the loss that measures the similarity between the two images. Many unsupervised image registration networks have been developed inspired by the STN.

One popular unsupervised image registration method, presented by Balakrishnan et al, involves a CNN-based approach, named VoxelMorph [19]. The model follows an encoder-decoder structure similar to U-Net, and a STN is incorporated at the end of the decoder to calculate the deformation field. They discovered that normalised cross-correlation as the similarity measure produced robust and reliable results. Additionally, to ensure smooth local spatial changes in the displacement field, a smoothness term was included in

the loss function.

However, the smoothness term alone has limited effectiveness in preventing folding of the displacement field, which can result in incorrect and non-diffeomorphic registration (invertible mapping). To address this issue, Zhang et al. proposed an inverse registration network [108]. They employed a fully convolutional network to align a pair of images (A and B) in both directions, generating two displacement fields, F_{AB} and F_{BA}. They minimised the difference between F_{AB} and its inverse field $-F_{AB}$, as well as between F_{BA} and its inverse field $-F_{BA}$, to ensure diffeomorphic registration. Additionally, they introduced an anti-folding loss to penalise folding pixels/voxels. The method achieved remarkable results on both Dice coefficients measured based on segmentation masks and the diffeomorphic properties.

In a different approach, Dalca et al. incorporated a vector integration layer into their FCN model [20], instead of using bidirectional image registration. They treated the output of the FCN as a stationary field and applied vector integration multiple times to obtain a diffeomorphic displacement field. By evaluating their method on a 3D brain dataset, they demonstrated similar Dice coefficient scores to other state-of-the-art methods, and with significantly improved diffeomorphic properties.

Based on this foundation, many other methods have been developed to improve unsupervised image registration. These methods include multi-scale structural models that use pyramidal structures [109] [110], adversarial-based method [111], vision transformer-based method [112], etc. These approaches have contributed to the ongoing progress in the field of unsupervised image registration.

2.4.3 Combination of Image Registration and Segmentation

The STN-based unsupervised image registration method has the capability to generate a displacement field as an intermediate variable. This allows for the estimation of segmentation in an unannotated image. Specifically, given a source image, a corresponding mask, and an unannotated target image, the displacement field can be estimated by inputting the paired source and target images into the unsupervised image registration model. The mask for the target image can then be obtained by applying the displacement to the mask of the source image. This approach was applied in VoxelMorph, where the Dice coefficient was used as an evaluation metric to assess the registration performance by comparing the segmented structures between the source and target images. Furthermore, an extension to VoxelMorph was to incorporate a segmentation loss to enhance the registration learning process [113]. The results demonstrated that the additional loss improved the Dice scores, indicating enhanced registration accuracy for structures with segmentation.

In addition to incorporating an additional segmentation loss function, another approach to address both image registration and segmentation tasks is through joint training of registration and segmentation networks. Qin et al. proposed a joint training model for motion estimation and segmentation using cardiac MRI data [114]. The unsupervised registration branch is employed to estimate the cardiac motion at different time points for the same patient, while the segmentation branch shared the same encoder to segment the cardiac structures in the corresponding time points. The displacement field generated by the registration branch is then utilised to warp the generated masks from an unannotated image to the ground truth masks of the target images. The model is optimised by minimising the similarity between the warped masks and the ground truths, effectively achieving a semi-supervised segmentation task. The results demonstrated the benefits of joint training for both the registration and

segmentation tasks.

Later, Xu et al. [21] introduced a more generic framework called DeepAtlas for weakly supervised registration and semi-supervised segmentation tasks. They combined the registration network and segmentation network, connecting them through an anatomy similarity loss. This loss measures the similarity between the segmentation of the target image and the warped segmentation of the source image. If either the source image or target image has a known mask, both models could be updated with the mask guidance. If not, the loss was set to 0, indicating that the models would not update themselves in that situation. By testing their model on two public 3D MRI datasets, they demonstrated significant improvements compared to using a single registration or segmentation network. Particularly noteworthy was the ability of DeepAtlas to achieve one-shot learning with remarkable performance by requiring only one annotated image.

Similarly, Mahapatra et al. conducted joint training of the segmentation and registration networks [115] and introduced a GAN to enhance the segmentation task. Their method demonstrated good performance on a breast X-ray dataset. Subsequently, Elmahdy et al. pursued a similar approach by incorporating an adversarial discriminator to evaluate the alignment quality of the registration branch [116]. They observed that the registration branch had a significant positive impact on the segmentation results. However, they found that the adversarial term primarily enhanced the performance of the registration branch and had limited influence on the segmentation branch. To the best of our knowledge, there has been limited research in the field of combining registration and segmentation methods.

2.5 Discussion and Conclusions

This chapter provides an overview of the background in image segmentation, semi-supervised learning, and image registration. For image segmentation,

classical mathematical-based methods such as region growing and grab cut are briefly introduced, followed by an overview of important deep learning-based methods including RNN-based, FCN-based, and encoder-decoder based approaches. Furthermore, a concise overview of image segmentation in the medical imaging domain reveals that encoder-decoder based methods are particularly well-suited for this field. However, supervised deep learning methods require a large amount of high-quality annotated data. Considering the limitations of manual image segmentation, an interactive image segmentation software is developed and described in chapter 3. This software, based on fully connected CRF, allows users to segment images using scribbles, enabling efficient segmentation of multiple masks for both 2D and 3D medical images. The software also includes an automatic recommendation feature for annotating the next slice in 3D, increasing efficiency and facilitating the collection of high-quality annotations from clinical experts.

Later, an introduction to semi-supervised learning was provided, highlighting three commonly used methods: GAN-based methods, consistency regularisation-based methods, and pseudo-labelling-based methods. These methods aim to extract valuable information from unannotated images to enhance the segmentation task. Among these approaches, self-learning and pseudo-labelling methods have shown to achieve state-of-the-art results. Additionally, ensemble techniques such as random forest, which combine multiple sub-models, have demonstrated significant contributions to the self-learning process. Inspired by these two methods, in chapter 4, an ensemble-based semi-supervised learning framework is proposed. By leveraging this framework and the interactive image segmentation software presented in chapter 3, we are able to develop a fully automated image segmentation model with reduced data annotation efforts from clinical experts.

In line with the objectives of this thesis, the final section of this chapter is dedicated to image registration. An overview of classical image registration methods is firstly provided, followed by the review of deep learning-based

approaches. Early deep learning methods for image registration utilised transformation matrices as the training target based on supervised learning. However, networks trained with simulated matrices often struggle when applied to real datasets, and collecting real transformation matrices can be challenging. The introduction of STN advanced the field, leading to the development of numerous unsupervised image registration methods based on STN principles. However, very few methods specifically addressed large-scale deformable image registration, which is a common scenario in medical image datasets. Moreover, STN-based methods necessitate constraints on the displacement field to prevent folding, but existing techniques either require additional training (e.g., bi-directional alignment) or increased computational complexity (e.g., anti-folding loss, vector integration, etc.). In chapter 5, a multi-scale diffeomorphic image registration method is proposed to address these limitations. Furthermore, a key feature of our proposed method is the ability to utilise masks to guide the alignment process, focusing specifically on selected regions of interest. This enhances the joint registration and segmentation framework in 6.

The final section of this chapter also provided an introduction to the combination of image registration and segmentation. While there is limited research in this area, it has been shown that joint training of the registration and segmentation networks can mutually benefit each other. A notable example is DeepAtlas, which has achieved impressive semi-supervised image segmentation results, even with just one annotated image. However, it is important to note that joint learning can be challenging, as one model's incorrect results can negatively affect the learning of both models, leading to a deterioration in performance over time. Therefore, in chapter 6, a novel quality assessment element is proposed to the joint image registration and segmentation framework. It incorporates an automatic evaluation mechanism to assess the quality of the segmentation results during iterative training. By gradually increasing the number of pseudo-masks used in training, it prevents the model from learning

incorrect information. This approach ensures a steady improvement in joint training, enhancing the overall performance of the model.

Chapter 3

Interactive Medical Image Segmentation

3.1 Introduction

In this chapter, the developed interactive image segmentation tool is introduced that provides efficient segmentation of multiple classes for both 2D and 3D medical images. The core segmentation method is based on a fast implementation of the fully connected conditional random field (CRF). The software also enables automatic recommendation of the next slice to be annotated in 3D, leading to a higher efficiency.

In summary, the key issues with the current interactive segmentation solutions for medical images are three fold. (1) Many image segmentation tools in computer vision work well in 2D images, but not many of them are applicable to 3D medical images, where the key barrier is computational time and effectiveness of user interactions in 3D. (2) Many solutions only focus on binary image segmentation, while multiple organs are often required to be segmented in medical images. (3) Many generic interactive image segmentation methods often work well in natural images with rich regional textures, which may not be directly applicable to medical images where multiple masks need to be assigned to image regions that have similar intensities. For example,

carpal bone segmentation in the wrist [117]. Moreover, different parameter settings are normally required for different images. This work aims to produce a generic medical image segmentation tool that works for both 2D and 3D images without any prior information or training process, and allows efficient user interactions.

To achieve the aim and address the aforementioned issues, the key contributions of this work are summarised as follows. (1) A fast CRF solver based on Gaussian approximation is adopted to achieve fast 2D and 3D image segmentation for multiple masks (up to 10 masks in the current implementation and can be easily extended.). (2) The software is featured with an automatic slice recommendation function to suggest the best slice to annotate, resulting in greatly improved image segmentation efficiency in 3D. (3) The method parameters have been optimally tuned, so that an "one size for all" setting is achieved, meaning no parameter adjustment is required for different medical image segmentation tasks. The developed tool has been evaluated on a variety of 2D and 3D medical image modalities and applications, in terms of segmentation accuracy, repeatability and computational time.

The remaining parts of this chapter is organised as follows: Section 3.2 presents the methodology underlying the proposed work. In section 3.3, the hyper-parameter configuration of the proposed method is presented, along with an overview of the software's user interface. Section 3.4 provides the evaluation results on various 2D and 3D medical image datasets. Finally, section 3.5 concludes the chapter by offering a summary and discussion.

3.2 Methodology

The goal of image segmentation is to annotate every pixel in the image with one of several predetermined object categories. In this work, it is formulated as maximum a posterior (MAP) inference in a CRF, which is defined over pixels in an image. The object classes (categories) are defined interactively by

the user using scribbles. The CRF is constructed by combining a smoothness term that maximises the annotation agreement between similar pixels and the likelihood of each pixel belonging to each of the user defined classes. The CRF is dynamically changed when more scribbles are added, leading to an iteratively refined segmentation result. The aim is to minimise the user interactions while achieving high quality image segmentation.

3.2.1 Fully Connected Conditional Random Field

In this section, the image segmentation task is formulated by CRF optimisation. For an image, the CRF model can be constructed as follows: The intensities of each pixel, denoted as $I = \{I_1, ..., I_N\}$, capture the pixel values across all N pixels in the image. The random field $X = \{X_1, ..., X_N\}$ indicates possible pixel-wise annotations. The domain of X is defined by a set of annotations, denoted as $L = 0, ..., K-1$, where K signifies the total number of annotation classes. For a binary class segmentation, $K = 2$ and $L = \{0, 1\}$. A configuration x represents a possible assignments of annotations for all the pixels in the image. The ground truth masks, represented as y, are the correct labels for the image. The goal is to make x as close to y as possible.

To achieve this task, Gibbs distribution are employed to estimate the likelihood of different annotation assignments. The Gibbs distribution is defined as below:

$$P(X|I) = \frac{1}{Z(I)} exp(-E(X|I)) \tag{3.1}$$

$$E(X|I) = \sum_{c \in C_G} \phi_c(X_c|I) \tag{3.2}$$

where $P(X|I)$ and $E(X|I)$ represent the probability and energy of annotation assignments X for a given image I, respectively. To understand how annotations are related to each other, a graph $G = (V, E)$ is created over the set of variables X. V and E represent vertices (individual pixels) and edges (connection between pixels) of the graph, respectively. The connected pixels are then

split into a set of cliques C_G. The potential functions ϕ_c are used to calculate the strength of the connections within each clique c. To ensure the annotation assignments are appropriately balanced, a partition function $Z(I)$ is used to normalise the distribution.

The Gibbs energy of a configuration x is $E(x|I) = \sum_{c \in C_G} \phi_c(x_c|I)$. The MAP method labels a random field of x^* that maximises $P(x|I)$. In a fully connected pairwise CRF model, G is a complete graph defined on X, meaning that every pair of pixels in the image is connected by an edge. C_G includes all unary (individual pixels) and pairwise (pairs of pixels) cliques. Therefore, the Gibbs free energy is expressed as:

$$E(x|I) = \sum_i \phi_u(x_i|I) + \sum_{i \neq j} \phi_p(x_i, x_j|I) \tag{3.3}$$

where i and j are the indices of pixels in I.

The unary term ϕ_u in Eq. (3.3) is normally computed independently for each pixel, indicating the probability of each pixel belongs to each of the classes. The pairwise potential ϕ_p represents the penalty of assigning labels to pixel i and j at the same time. In fully connected CRF model, the pairwise cliques describe all two pairs of random variables. Subsequently, the mean field theory can be employed to produce an asymptotic solution.

In the implementation, the same pairwise cost ϕ_p as proposed by Krähenbühl et al. [33] is used. However, a different unary term ϕ_u is used, which is described in the next section. The pairwise cost consists of two terms that model the appearance and smoothness between pairs of pixels, expressed as:

$$\phi_p(x_i, x_j) = [x_i \neq x_j] g(i, j) \tag{3.4}$$

$$g(i,j) = \omega_1 \exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}) + \omega_2 \exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}) \tag{3.5}$$

In Eq. (3.4), $[x_i \neq x_j]$ is an indicator function that indicates if the two labels at pixel i and j are the same. The first term (i.e. appearance term) in

$g(i, j)$ encourages nearby pixels (determined by pixel location p) with similar intensities (denoted as I) to be the same class. The degrees of nearness and similarity are controlled by the parameters θ_α and θ_β respectively. The second term (i.e. smoothness) helps in removing small isolated regions that is controlled by θ_γ. ω_1 and ω_2 are the weights to balance the two terms. The parameters are determined experimentally using many different modalities of medical images and reported in section 3.3.1.

3.2.2 Unary Term for Interactive Image Segmentation

The unary term could be modelled by generative models, such as using histogram or mixture Gaussian to model the data distributions of different classes. However, these methods normally require sufficient number of samples to learn. With limited scribbles of each class, especially in the first few iterations, the use of generative model is neither accurate nor computationally efficient. Additionally, unlike other machine learning (include deep learning) based unary term modelling methods (e.g. [118] [119]), the proposed method does not require multiple images of the same object and a pre-training step. In the proposed method, the unary term in Eq. (3.3) is designed by considering both the intensity similarity (the first term in Eq. (3.6)) and distance (the second term in Eq. (3.6)) of a pixel to the scribbles annotated by the user in a simple Gaussian weighted manner, which is expressed as below.

$$\phi_u^l(i) = \lambda \exp(-0.5(\frac{I_i - m^l}{\sigma_1^l})^2) + (1-\lambda)\exp(-0.5(\frac{d_i^l}{\sigma_2})^2) \tag{3.6}$$

where m^l and σ_1^l are the mean and standard deviation of the intensity values annotated by the user for the l^{th} class respectively. They are calculated and updated based on the annotated pixels during the user interaction process. The first term measures the likelihood of a pixel i belonging to class l, resulting in a probability map P_l. Fig. 3.1 (a) and (b) show an example image and some user annotations (yellow: foreground; white: background) respectively.

Fig. 3.1 (c) and (d) are the intensity-based probability maps P_0 and P_1 for the background and the foreground respectively. A brighter pixel indicates a higher probability of belonging to the corresponding class. d_i^l in Eq. (3.6) is the length of the minimum path between the i^{th} pixel to the nearest labelled pixel of the l^{th} class, which is calculated as geodesic distance. When calculating the minimum path, the locations along the path are weighted by the gradient of the probability image P_l. Therefore, the minimum path is the route that generates the smallest changes of P_l between two locations. The implementation is based on geodesic time algorithm [120]. The capture range of the distance measure is controlled by the parameter σ_2. Different from σ_1, σ_2 is predefined as a hyper-parameter in table (3.1). Then the Gaussian weighted geodesic distance (second term in Eq. (3.6)) is computed, as shown in Fig. 3.1 (e) (background) and Fig. 3.1 (f) (foreground) respectively. λ is used to balance the intensity term and the distance term. The parameter settings are discussed in section 3.3.1.

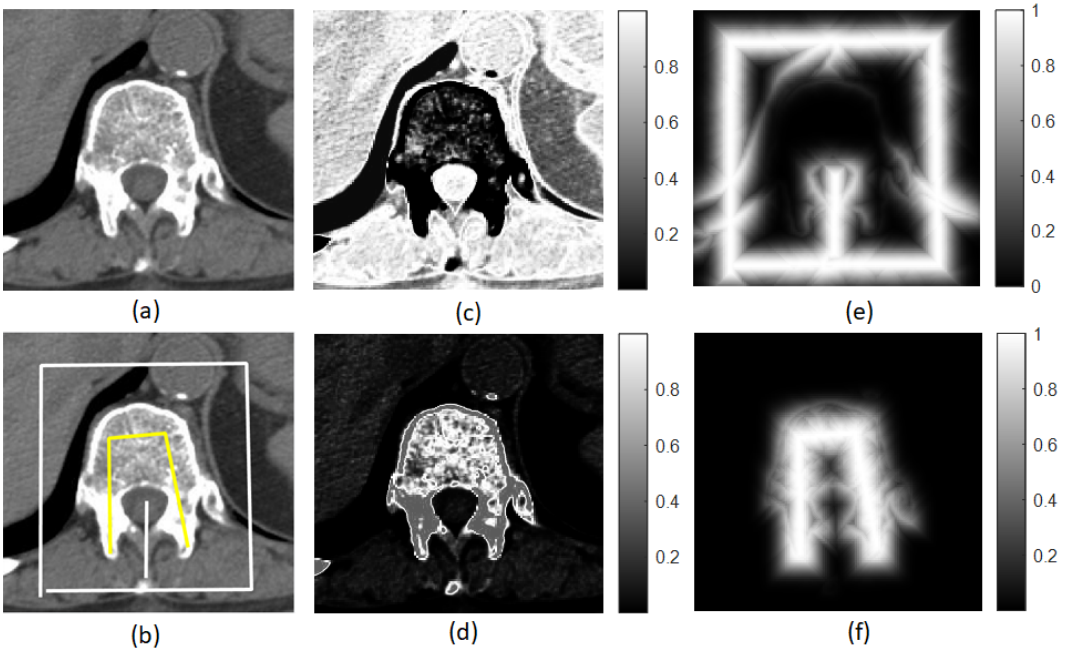

Figure 3.1: *(a) An example image. (b) User annotation: white-background ($l = 0$); yellow-foreground ($l = 1$). (c) and (d) are the probability maps of background (P_0) and foreground (P_1) respectively. (e) and (f) are, respectively, the Gaussian weighted geodesic distance maps of background and foreground (second term in Eq. (3.6)) with $\sigma_2 = 10$ pixels.*

3.2.3 Image Segmentation and Refinement

The above constructed CRF can be solved by mean field theory. Krähenbühl et. al. [33] developed an iterative filter-based method for performing fast approximate maximum posterior marginal inference. The number of iterations is denoted as t. This fully connected CRF optimisation method has not been applied to interactive image segmentation previously.

In the user annotation process, different class labels are required to be assigned to different objects of interest. The correct number of class labels is not required in the first annotation step and more class labels can be added at any stage of the user interaction. The annotation can be corrected/overwritten by new labels at the same location. For adjacent objects that share similar intensities, some scribbles of each class are expected to be drawn close to the shared boundary. For 3D volume annotation, annotations from different views are conducted separately but recorded in a single 3D volume. When a voxel is assigned by multiple labels from different views, the latest assigned label is used. For 3D images, the fully connected CRF optimisation is performed in 3D.

Fig. 3.2 (a) shows an example image and Fig. 3.2 (b) is the scribbles that a user annotated. Note that these scribbles were interactively added based on intermediate segmentation results. The final result based on the annotations in Fig. 3.2 (b) is shown in Fig. 3.2 (c). As indicated by the white arrows in Fig. 3.2 (c), there are holes in certain regions and some inaccurately segmented regions along the boundaries of some organs. The user can keep adding more scribbles to these inaccurate regions until satisfied. However, it could be a tedious work to accurately refine these boundaries manually. Alternatively, an automatic segmentation refinement step was added to help the user in reducing the number of interactions.

The segmentation refinement is achieved as follows. (1) Replacing the probability map G_c (first term in Eq. (3.6)) by the output probability map

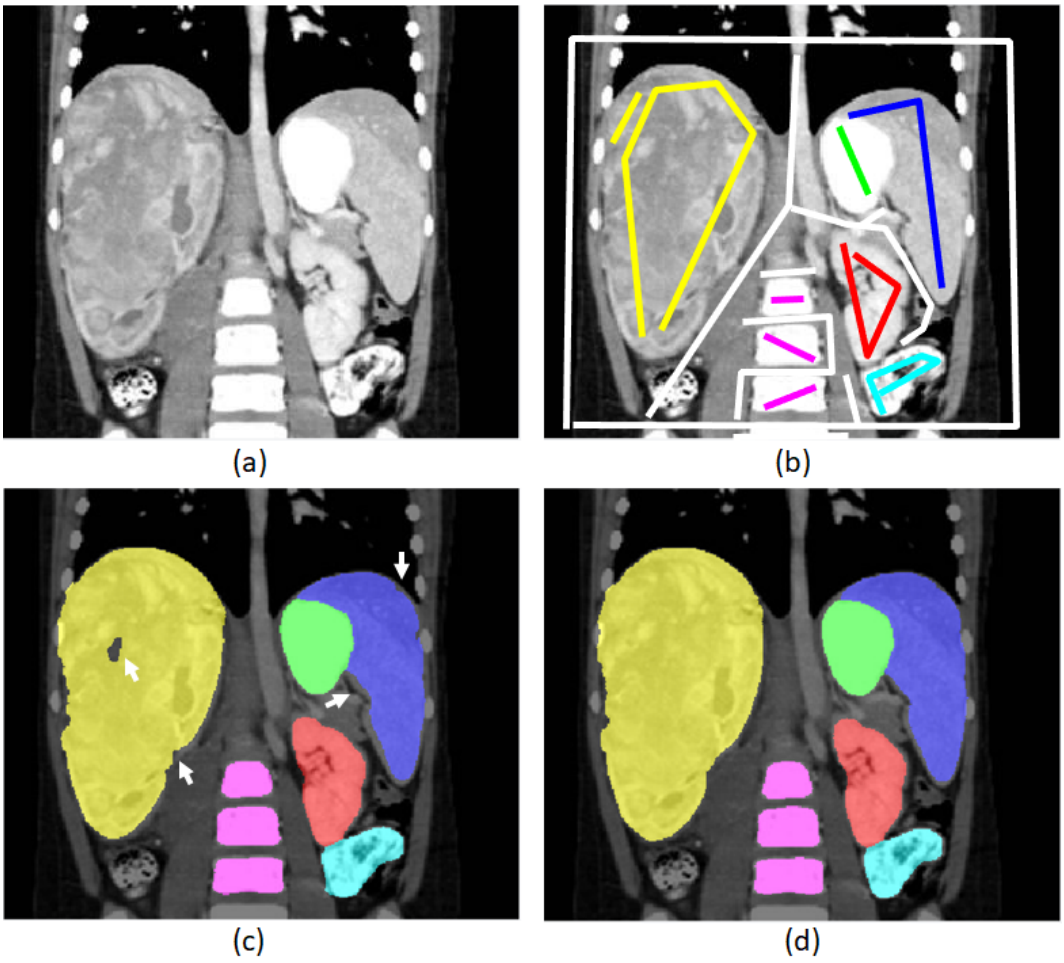

Figure 3.2: *(a) An example image. (b) User annotation with multiple classes. (c) Segmentation result before refinement with some inaccurate regions indicated by white arrows. (d) Segmentation result after automatic refinement.*

from the current CRF solution. (2) The geodesic distance map d^c in Eq. (3.6) is then recalculated based on the new G_c. (3) The CRF optimisation is applied again using the updated unary term, leading to filled holes and refined boundaries as shown in Fig. 3.2 (d). In the implementation, this refinement step runs as an extra step on every intermediate result, not only on the final result. It requires slightly longer computational time, but leads to fewer user interactions.

3.2.4 Entropy-based Slice Recommendation

For 3D image segmentation, the 3D volume normally contains many slices (typically more than one hundred) in the three different views (i.e. axial, coro-

nal and sagittal). In this interactive image segmentation tool, the user is asked to start the annotation from the middle slice in each of the three views, which is more likely to contain the object of interest. By only annotating one slice of any or all of the three views, the tool is able to generate an initial segmentation result. For other slices which are not annotated by the user, the closer to the annotated slice the more accurate the segmentation will be. Typically, the user needs to validate the segmentation result in each slice of all views and correct the results if not satisfied, which is a time-consuming process. In this software, the next slice in each of the three views that potentially contains the largest segmentation error is automatically suggested, based on the calculation of entropy. Entropy has been used as an indication of uncertainty by many research works (e.g Gal [121]). Hoebel et. al. have reported high correlations between the uncertainty measures and the corresponding Dice coefficient values [122]. In this case, the entropy H for the 3D volume is calculated as below.

$$H = -\sum_{c=0}^{K-1} G_c \log(G_c) \tag{3.7}$$

where G_c is the probability map for the c^{th} class generated by the CRF optimisation. K is the total number of classes. Subsequently, the average entropy value for each slice in each of the three views is calculated. A larger entropy value indicates a higher uncertainty of the segmentation result. As the example shown in Fig. 3.3, higher entropy values are found at boundaries of the segmented objects that corresponding to larger segmentation errors (also see the highlighted regions by the white arrows in the middle image). The slice that produces the largest average entropy value for each view is suggested to the user for further annotation. The experiments also show that this slice recommendation function significantly improved the annotation efficiency, especially in the first few interactive actions (see section 3.4.5). This function is an optional feature to the user, where a button in the GUI has to be clicked every time of requiring a suggestion. Note that the user still needs to validate

the segmentation result on every slice in each view to ensure a high quality segmentation output.

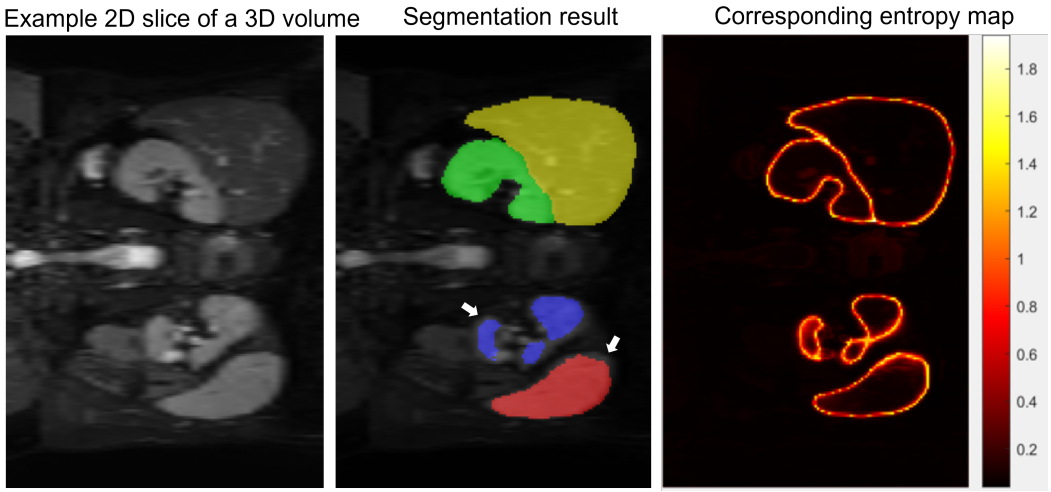

Figure 3.3: *Images from left to right are an 2D slice of a 3D volume in the CHAOS dataset, the segmentation result at an intermediate iteration and the corresponding entropy map. White arrows indicate the image regions with larger segmentation errors that correspond to larger entropy values.*

3.3 Parameter Settings and Graphical User Interface

3.3.1 Parameter Settings

The meaning and values for all the hyper parameters described in section 3.2 are listed in table 3.1. The parameter values were determined by evaluating the software on various 2D and 3D images described in section 3.4.1. In the context of medical image segmentation, the parameters were optimised in favour of responding to the user's annotations to improve the segmentation accuracy rather than minimising the number of user's interactions. Hence relatively smaller values were used for σ_2 and λ in Eq. (3.6) to make the tool more responsive to the user's annotations. After tuning the parameters, the values were all fixed for all testing experiments reported in this chapter. As shown in the graphical user interface in Fig. 3.4, the proposed software does not require the user to adjust any parameters.

Table 3.1: Hyper parameter setting.

Symbol	Meaning	Value
θ_α	Nearness controller for the appearance kernel (Eq.(3.5))	20 (2D & 3D)
θ_β	Similarity controller for the appearance kernel (Eq.(3.5))	1 (2D & 3D)
θ_γ	Controller for the smoothness kernel (Eq. (3.5))	1 (2D & 3D)
ω_1	Weight of the Gaussian appearance kernel (Eq. (3.5))	1 (2D & 3D)
ω_2	Weight of the smoothness kernel (Eq. (3.5))	5 (2D & 3D)
σ_2	Distance measure controller (Eq. (3.6))	4 (2D & 3D)
λ	Weight for balancing the intensity term and the distance term (Eq. (3.6))	0.1 (2D & 3D)
t	Number of iterations in CRF optimisation	10 (2D);3 (3D)

3.3.2 Graphical User Interface

The software was implemented in Matlab (version 2020b) and compiled to an executable file (.exe). The core functions of CRF optimisation were written in C++. Currently, it only supports Windows operation system and has been tested on Windows 10. However, the source code is also freely available for research purpose which can be compiled in other operating systems.

The graphical user interface is shown in Fig. 3.4. The basic functions labelled in the figure are briefly described as follows. (1) Load a 2D image to be segmented with the supporting file formats of .jpg, .tiff, .bmp, .png, DICOM and .mat. (2) Load a 3D volume (or multiple 2D slices) with the supporting formats of .mat, DICOM and .nii. (3) Load segmentation result in .mat or .nii format. (4) Re-sample the 3D volume into isotropic physical unit (mm). (5) Perform CRF segmentation after user annotation. (6) Enable/disable overlapping the segmented results to the original image. (7) Automatically suggest the best slice to be annotated for each of the views in 3D. (8) Display the measured areas (2D) or volume (3D) in physical unit (if unit is known) for each of the segmented classes. (9) Save the segmented result in the supported formats:

Figure 3.4: *Graphical user interface of the developed software. An example 3D wrist MRI, in which 10 bones were segmented by the proposed software as indicated by different colours.*

2D (.mat, .png, .bmp and .tiff) and 3D (.mat and .nii). (10) Clear all views and data. Reset all parameters to the default settings. (11) Exit the software and clear memory. (12) Information bar that indicates the current status of calculation. (13) Brief information about the software and user instructions. (14) Slide bars to change the slices in the corresponding views for 3D volumes. (15) Crop a smaller region of interest from the input 2D or 3D image in the corresponding view. (16) A drop down list of labels in the corresponding view for user annotation. Currently it supports up to 10 foreground classes plus the background. (17) Enable/disable 3D mesh view of the segmented result for a 3D volume. (18) Tool for manipulation of the 3D mesh model. (19) Image viewer for displaying and annotating 2D and 3D images. 2D image is displayed in the top left viewing window. (20) Viewing window for displaying the 3D mesh model of the segmentation result for a 3D volume.

3.4 Method Evaluation

3.4.1 Materials

Several medical imaging datasets were used for method evaluation.

Combined (CT-MR) Healthy Abdominal Organ Segmentation dataset (CHAOS) [123]: the CHAOS challenge aims to segment liver, kidneys and spleen in the abdominal region in CT and MRI data. The manual annotation process was produced by two teams, both of which include a radiology expert and an experienced medical image processing scientist. After the manual annotation of both teams, a third radiology expert and another medical imaging scientist analysed the labels, which were fine-tuned according to the discussions between annotators and controllers. The CHAOS dataset has high quality ground truth segmentation masks, hence selected to evaluate the segmentation accuracy of the proposed tool. The CT dataset was acquired from 40 different patients and only has the segmentation mask of liver. The MRI dataset contains two sequences (i.e. T1- DUAL and T2-SPIR) of 40 patients from 1.5T MRI scanners. The segmentation of T2-SPIR MRI dataset contains liver, both kidneys and spleen, which is a more challenging multi-label segmentation task, hence selected as the dataset to evaluate the proposed method.

Wrist CT dataset: this wrist CT dataset was used in [124] that contains CT image from 25 subjects. Each subject was imaged at five different wrist poses: neutral and four extreme poses in radial-ulnar and flexion-extension. The pixel spacing is $0.29mm \times 0.29mm$ with a slice thickness of $0.625mm$. It is a challenging image segmentation task, as 10 bones (i.e. ulnar, radius and eight carpal bones) are required to be segmented in a small wrist region of the CT image. All the bones have similar intensity values and in close contact with each other.

For the purpose of parameter tuning as described in section 3.3.1, a variety of medical images that cover different imaging modalities and different organs were obtained from a number of local and public datasets, such as plain wrist

X-ray image from Chen et al. [125], contrast enhanced breast MRI from [126], ultrasound breast image from Al-Dhabyani et al. [127], retinal images from Budai et al. [128]. These images were used as a validation set for parameter tuning of the proposed method. Some qualitative segmentation results of these images are presented in section 3.4.6.

3.4.2 Comparison of Local Pair-wise CRF and Fully-connected CRF

One of the most widely used CRF based optimisation that has been applied to interactive image segmentation is based on the pair-wise potential of nearest neighbours [34] (denoted as PwCRF). The performance of PwCRF and the fully-connected CRF (denoted as FcCRF) was compared, in terms of qualitative segmentation accuracy and computational time using the same user annotations.

The PwCRF implementation is the baseline method implemented by Kohli et al. [129]. The core PwCRF optimisation using α expansion was implemented in C++ based on the paper [32]. The core FcCRF optimisation using mean field approximation was adapted from the C++ implementation by Kamnitsas et al. [130]. The evaluation was performed on the same machine.

The comparison based on a binary class segmentation and a multi-class segmentation was performed. Images in Fig. 3.5 show the original input images, user annotations and segmentation results using PwCRF and FcCRF respectively. It can be seen that the segmentation result of using PwCRF, especially for the multi-class case, is less accurate at the boundaries of the objects than the FcCRF method (highlighted by white arrows).

Additionally, using the bottom image of Fig. 3.5, The CRF optimisation time by varying the number of segmentation classes (i.e. 2, 4, 6 and 7 classes including background) was compared. As reported in table 3.2, the FcCRF optimisation time is not highly dependent on the complexity of the annota-

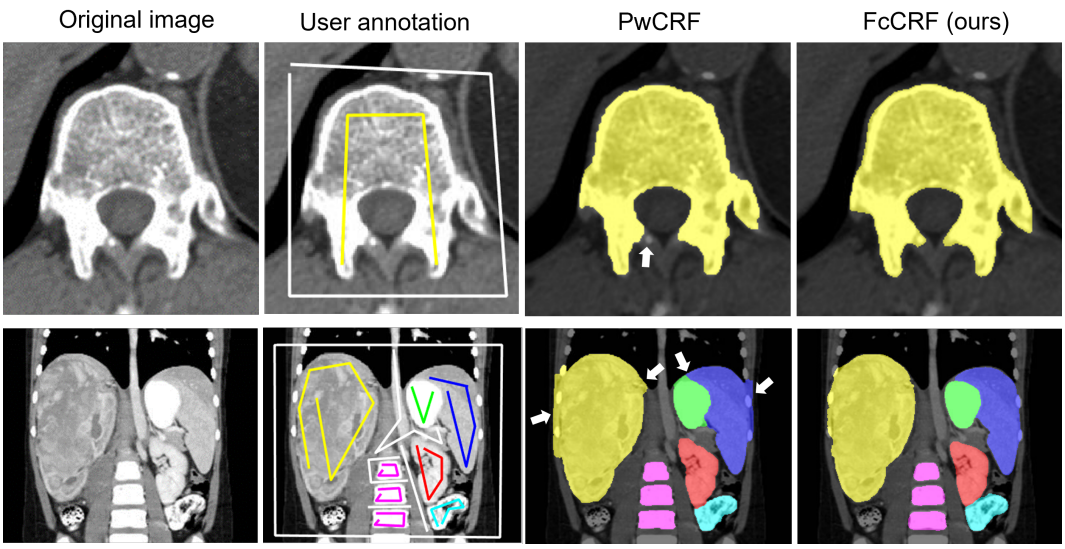

Figure 3.5: *Comparison of local pair-wise CRF (PwCRF) and fully-connected CRF (FcCRF). Top row: an example of binary class segmentation. Bottom row: an example of multiple class segmentation. White arrows indicate inaccurate segmentation locations.*

tion and the number of classes. However, the computational time of PwCRF using α expansion is highly dependent on the number of classes that increases significantly when the number of classes increases. Hence the total run time of FcCRF is much shorter than the PwCRF, especially for a multi-class scenario, due to a more accurate segmentation result at each iteration (therefore fewer number of interactions) and shorter optimisation time at each iteration. This is the main reason that FcCRF is selected in this interactive image segmentation software.

Table 3.2: Computational time using local pair-wise CRF (PwCRF) and fully-connected CRF (FcCRF) for segmenting different number of classes for the bottom image in Fig. 3.5. The time reported is only for CRF optimisation for a fair comparison.

Method	Binary class	4 Classes	6 Classes	7 Classes
PwCRF	0.15 s	0.84 s	1.56 s	2.24 s
FcCRF	0.41 s	0.47 s	0.63 s	0.65 s

3.4.3 Evaluation on Segmentation Accuracy

Five T2-SPIR MRIs from the CHAOS dataset were randomly selected for evaluating the segmentation accuracy of a multi-label segmentation task in 3D

image. Dice coefficient (DC) and average symmetric surface distance (ASSD) were used as the evaluation metrics. The DC is a widely used measurement in image segmentation evaluation, which indicates the volume agreement between the generated segmentation result and the ground truth segmentation mask (i.e. 0 and 1 indicate the worst and best segmentation results respectively). The ASSD determines the average difference between the surface of the segmented object and the ground truth segmentation mask in 3D. After the border voxels of the segmentation output and the ground truth mask are determined, those voxels that have at least one neighbour from a predefined neighbourhood that does not belong to the object are collected. For each collected voxel, the closest voxel in the other set is determined and the average of all these distances derive the ASSD measure (0 mm for a perfect segmentation, max distance of the image for the worst case). The mean $\pm$ standard deviation values of DC and ASSD for the five MRIs are reported in table 3.3.

For interactive image segmentation, the segmentation result is highly dependent on the number of interactions and experience of the annotator. The results presented in table 3.3 were produced by an annotator without medical background. In the proposed method, the volume was firstly cropped and resampled to an isotropic volume, and the segmentation was then performed in 3D. The size of the volume was approximately $120 \times 170 \times 120$ (the sizes of different volumes are slightly different) with the voxel size of $2mm^3$. The software ran on a laptop with an Intel i5-6300U 2.4 GHz processor and 8 GB memory. The total segmentation time is highly dependent on the size of the volume and the number of interactions required. The average number of interactions and time in these experiments were approximately 20 and 25 minutes per volume respectively. The segmentation accuracy is comparable to the results reported in the literature [131].

Fig. 3.6 shows the visual result of an example that produced the lowest mean DC value of the four organs (liver: 0.88; both kidneys: 0.86; spleen: 0.83) in the five segmented volumes. From the 2D slices in Fig. 3.6, it can be

Table 3.3: Segmentation accuracy of CHAOS dataset for liver, kidney and spleen segmentation. Mean ± standard deviation values of Dice Coefficient (DC) and Average Symmetric Surface Distance (ASSD) are reported.

Metrics	Liver	Left kidney	Right kidney	Spleen
DC	0.923 ± 0.002	0.906 ± 0.023	0.894 ± 0.018	0.875 ± 0.019
ASSD(mm)	2.504 ± 0.080	1.616 ± 0.388	1.752 ± 0.298	1.623 ± 0.248

seen that the segmentation result using the proposed software (green contours) is agreed with the ground truth annotation (red contours) in most regions, except for the peripheral regions of some organs (e.g. pelvis of kidney). This disagreement is highly related to the experience and knowledge of the annotator. Technically, higher segmentation accuracy can be achieved by more user interactions to refine these regions. From the mesh models shown in Fig. 3.6, it can be seen that the ground truth annotation is performed in a multiple 2D slice manner, while the proposed method is performed in 3D, which may also lead to the result discrepancy.

3.4.4 Evaluation on Repeatability and Reliability

To evaluate the repeatability of the interactive image segmentation tool, three annotators performed image segmentation on the same set of three T2-SPIR MRIs from the CHAOS dataset. Prior to the experiments, the three annotators were briefly trained by showing the organs of interest and the corresponding reference annotation of an independent T2-SPIR image. This helps to minimise the effects of knowledge discrepancy between the annotators. The annotators were given sufficient time to complete the segmentation task until they were satisfied with the segmentation result. Subsequently, the intra-class correlation coefficient (ICC) [132] was calculated based on the DC values to measure the performance consistency of different annotators. The average ICC score for all class labels is 0.7618, which indicates a good agreement (follow the guideline by Koo et al. [133]) between the segmentation results of different annotators using the proposed software.

One of the key aims of medical image segmentation is the quantitative

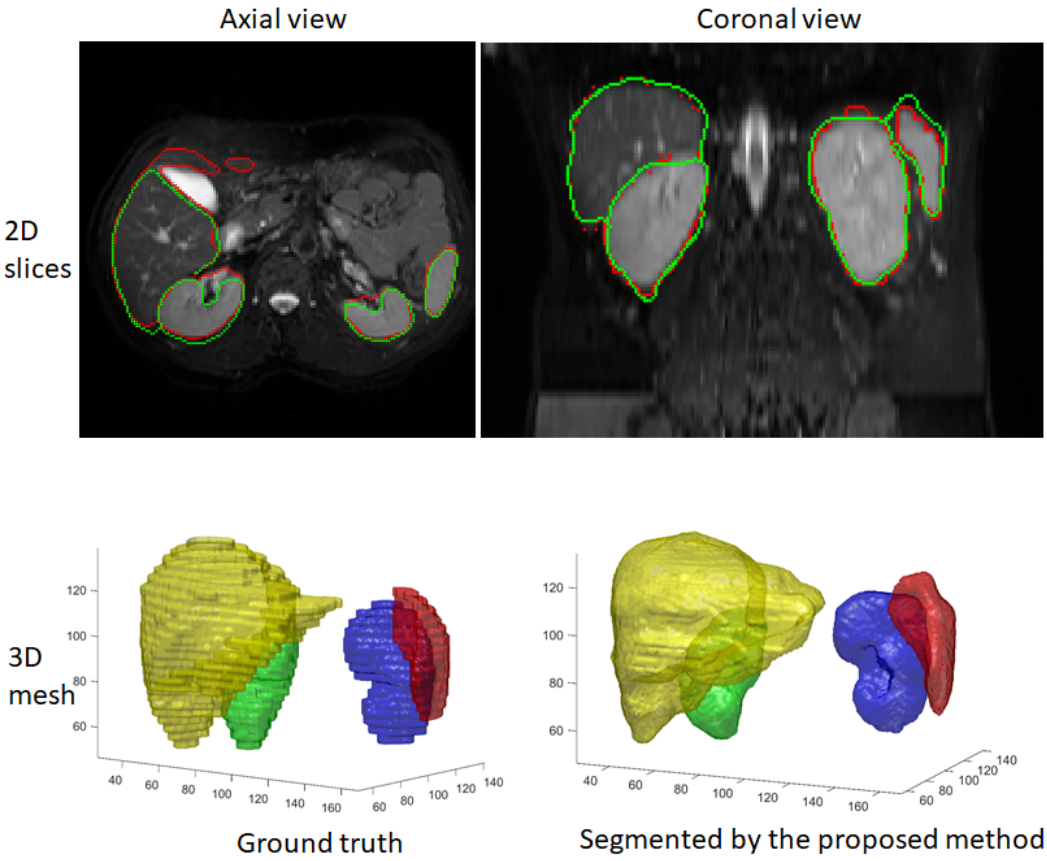

Figure 3.6: *Top row: visual segmentation result in axial view and coronal view (red: ground truth; green: the proposed method). Bottom row: visual segmentation results in 3D mesh model of ground truth and segmentation result using the proposed method (liver: yellow; kidneys: green & blue; spleen: red).*

measurement of the region of interest, such as volumes of organs or tumours. The reliability of the measured volume is crucial for downstream clinical decision making tasks. Here, the wrist CT dataset was used to assess the reliability of the segmentation result produced by the proposed software. A single annotator performed segmentation of eight carpal bones using the proposed software on the CT images of three subjects, each contains CT volumes that were captured at five different wrist poses. The ulnar and radius bones were also segmented as the reference bones but were not considered for the performance evaluation. The assumption is that the measured volumes of the carpal bones at different wrist poses of the same subject should be the same.

Fig. 3.7 shows the results of the carpal bone segmentation in different wrist poses of the same subject using the proposed software. The variations of the measured bone volumes are listed in table 3.4. The Std% value in table

3.4 is calculated by using the standard deviation of the bone volumes measured from different poses divided by the corresponding mean bone volumes, and then averaged across the three subjects. An average of 2-3% of volume variations in the segmentation results indicates a small error range in measuring the bone volumes. The ICC was also calculated based on the segmented bone volumes to measure the consistency of the measurements across different poses. An ICC value of 0.9769 is achieved, which indicates a high consistency of the bone volumes measured in different wrist poses.

Table 3.4: Percentage of standard deviation for volume of carpal bones segmented from different poses of the same subject. The carpal bones are: Triquetrum (Tri), Lunate (Lun), Scaphoid (Sca), Pisiform (Pis), Hamate (Ham), Capitate(Cap), Trapezoid (Trd) and Trapezium (Trm).

Bones	Tri	Lun	Sca	Pis	Ham	Cap	Trd	Trm
Std %	2.69	3.05	2.99	2.81	2.11	3.15	1.41	2.41

3.4.5 Evaluation on Efficiency

The high computational efficiency of the fully-connected CRF solution used in this method has been demonstrated and compared to other CRF optimisation methods by Krähenbühl et al. [33].

In this section, the efficiency of the proposed slice recommendation function is firstly demonstrated. One randomly selected T2-SPIR MRI in the CHAOS dataset was used in this experiment. Two annotators were asked to segment four organs (liver, two kidneys and spleen) from the same 3D volume with the same given initial annotation. Each of the annotators performed twice on the segmentation task, one with and one without using the slice recommendation function. “With the slice recommendation” was performed first, hence the result of “without slice recommendation” could be slightly better than reality due to the previous familiarisation with the data. Despite this potential bias, the segmentation quality measured by Dice coefficient (average of the four organs) from both annotators increased much faster by using the slice recommendation function, especially in the first few user interactions (shown

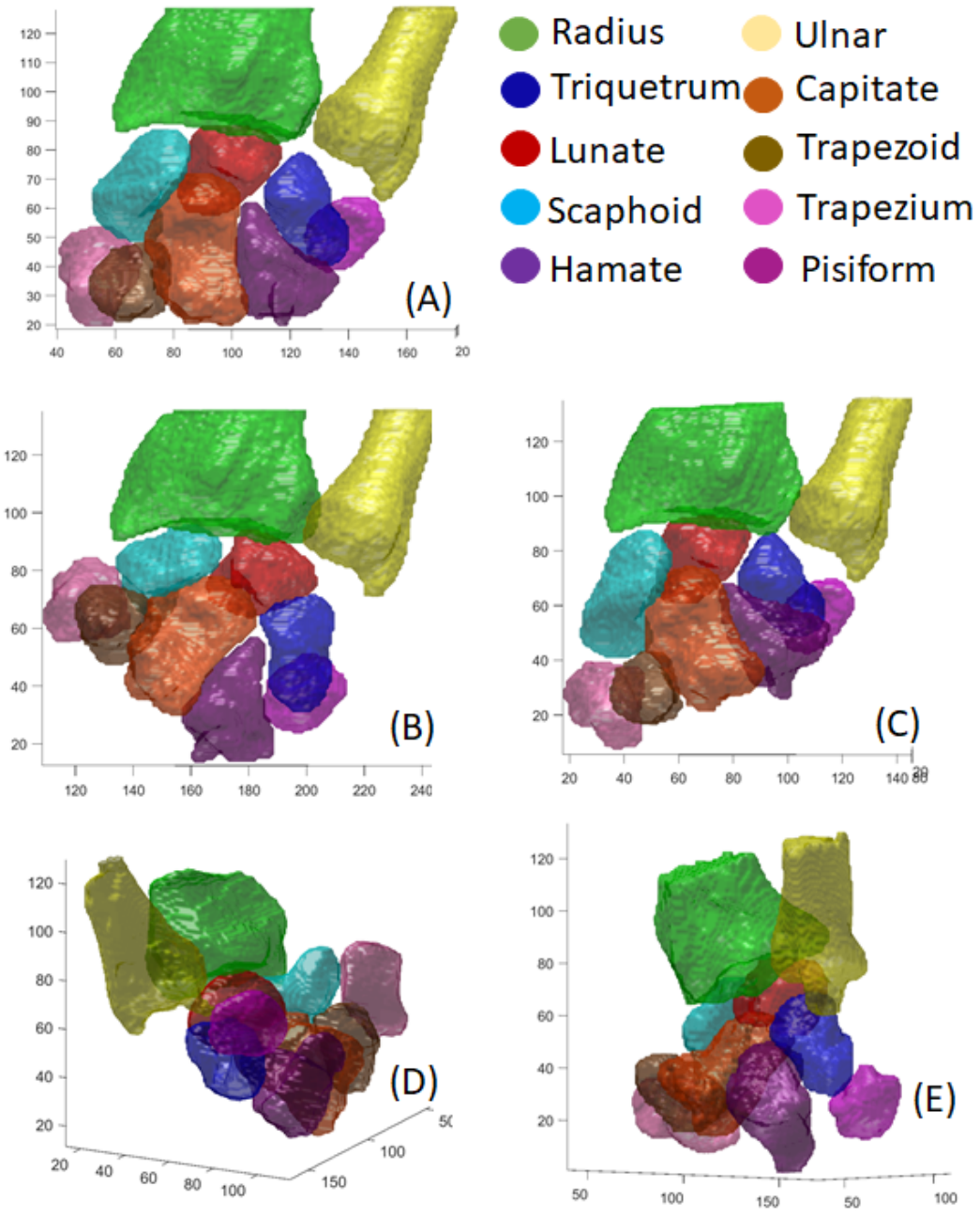

Figure 3.7: *Segmentation results of carpal bones in CT volumes of different wrist poses from the same subject. (A) Neutral (B) Radial-deviation (C) Ulnar-deviation (D) Flexion (E) Extension.*

in Fig. 3.8). This demonstrates the improved segmentation efficiency by using the proposed slice recommendation function.

Next, the segmentation efficiency of the developed software is compared to a widely used manual segmentation tool (i.e. 3D Slicer [11]) by drawing polynomial lines in a slice by slice manner. A single annotator performed organ segmentation using 3D Slicer and the proposed software on a T2-SPIR

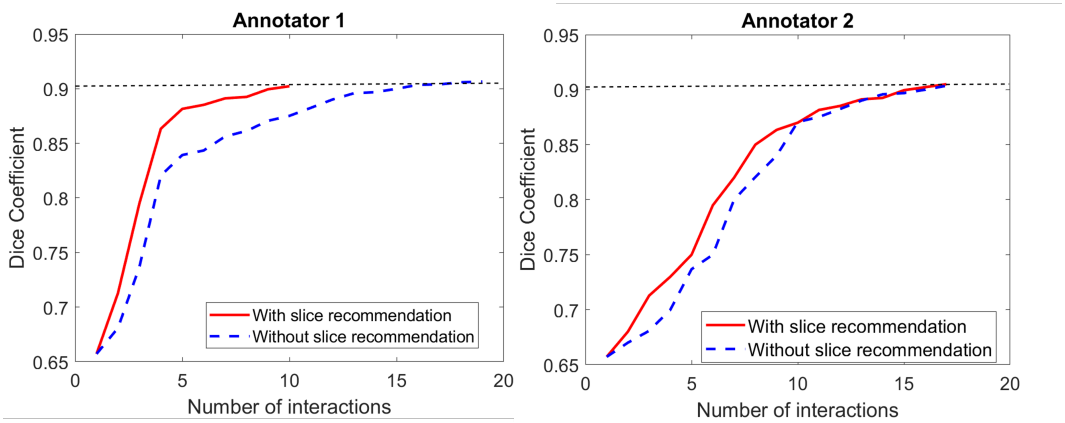

Figure 3.8: *Comparison of segmentation quality (Dice coefficient) with/without slice recommendation from two annotators.*

MRI in the CHAOS dataset. The DC value using 3D slicer for liver, left kidney, right kidney, and spleen were 0.91, 0.91, 0.89 and 0.88, which achieved similar performance to the proposed method (refer to table 3.3). However, the time used in 3D slicer was 50 minutes, which was significantly higher than the time used in the proposed method (average of 25 minutes).

It is even more challenging in segmenting the wrist CT data. Due to multiple carpal bones having similar shapes and intensities, it is extremely difficult to maintain the track of the class labels of different bones during the slice by slice manual segmentation process. Using 3D slicer, the annotator made a lot of efforts in checking the slice-wise context to ensure consistent class labels across different slices. In contrast, the proposed method is performed in 3D and the whole 3D volume is labelled by only annotating a few slices in different views. For the wrist CT data, 3D slicer required about 90 minutes to label a single volume, and the annotator needs to have a good knowledge about the anatomy of the wrist. The software only took about 30 minutes to segment the carpal bones. Especially, the proposed software does not require accurate contour tracing which needs lower concentration level from the annotator.

3.4.6 Qualitative Segmentation Results

Besides the above systematic evaluations, some qualitative segmentation results of using the proposed software on a variety of medical images are presented in Fig. 3.9, including plain X-ray images for bone segmentation, 3D contrast enhanced breast MRI for tumour segmentation, breast ultrasound image for tumour segmentation and retinal image for blood vessel, macula and optical disc segmentation. It can be observed from Fig. 3.9 that the software works well for segmenting organs, tumours in any of the given medical modalities, but failed to segment part of the blood vessels in the retinal image. This type of thin linear structure, which distributed across the whole image, requires the user annotations to cover the whole image. In this case, it makes the user annotation extremely challenging and time consuming using the proposed method. In this example, the user only annotated a very small portion of the image leading to an unsatisfactory result for part of the blood vessels. Hence, alternative solutions are strongly recommended if such a linear structure needs to be segmented. For example, a deep learning method that handles inaccurate annotations were proposed by Zhang et al. [59] for segmenting linear structures.

3.5 Discussion and Conclusions

In this chapter, an efficient software using fully connected CRF optimiser to achieve multi-class 2D and 3D medical image segmentation is presented. Based on the CRF optimiser, an interactive image segmentation tool for medical image analysis is developed. This tool does not require parameter tuning for different image modalities and dimensions. It is also featured with a slice recommendation function to achieve efficient user interactions for 3D images. The method has been comprehensively evaluated in terms of segmentation accuracy, repeatability, reliability and efficiency on different medical imaging datasets and applications. This method performs well on the segmentation

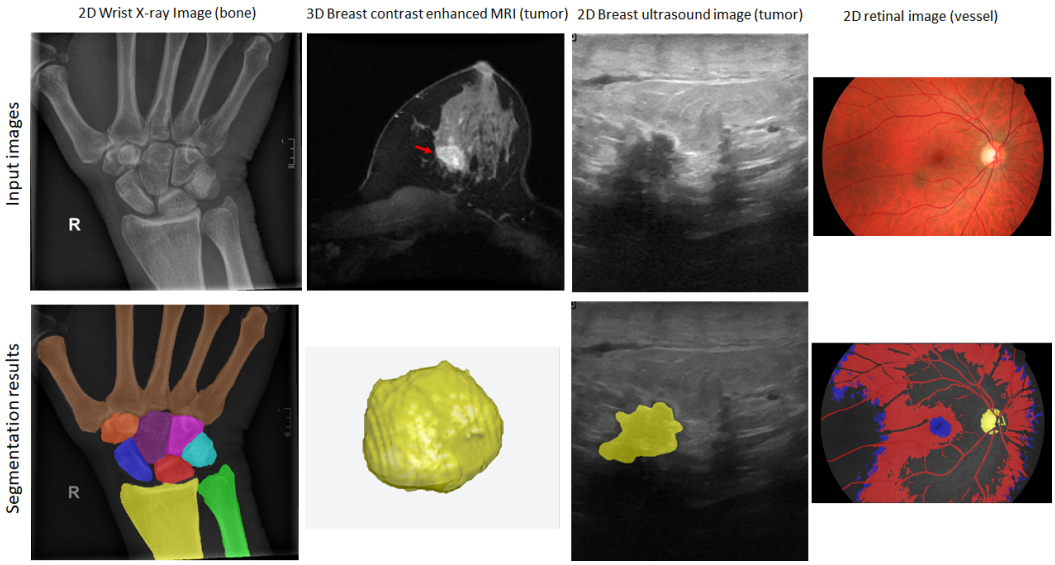

Figure 3.9: *Qualitative segmentation results of different medical images and applications. The bones, organs and tumours can be efficiently segmented by the proposed method. For the retinal image segmentation, it requires tremendous user annotation efforts to segment the linear structures in the whole image. Hence, it is not recommend to use the proposed method in segmenting thin linear structures.*

of regular shaped objects (e.g. organs, bones, tumours, etc.), but it is less efficient in segmenting thin linear structures, such as blood vessels in retinal image.

The software is freely available for research purposes. It provides an effective way to annotate the images in a given new dataset. However, even though doctors can efficiently and accurately label images using this tool, labelling a large number of images still requires significant efforts and time. Hence, in the next chapter, semi-supervised learning method for image segmentation is explored. This technique allows a machine learning model to achieve automatic image segmentation by learning from a dataset that contains only a small number of annotated data.

Chapter 4

Semi-supervised Image Segmentation Using Model Ensemble

4.1 Introduction

In the previous chapter, a highly efficient interactive image segmentation tool was developed, providing an approach to obtain annotated data from human experts. This tool facilitates the process of acquiring a limited amount of annotated data from experts when confronted with a new dataset. By combining this annotated dataset with a larger amount of unannotated data, a semi-supervised learning approach can be used to train an automatic segmentation model.

To establish a generic semi-supervised learning approach for image segmentation, the decision has been made to employ a pseudo-labelling method that can be easily applied to diverse data types. As outlined in section 2.3.3, when applied to image segmentation tasks, the pseudo-labelling method requires a strategy to ensure the accuracy of the pseudo segmented mask.

Given this requirement, ensemble learning, a widely used technique in machine learning [134], was adopted. It is an effective method for rectifying errors

in pseudo predictions. The basic idea of ensemble learning is to train multiple weak decision makers and then combine them to generate a final decision. Several classical classifiers, including random forests [135], have successfully employed this idea to achieve state-of-the-art performance prior to the advent of deep learning methods. A notable example of the benefits of ensemble learning can be observed in the work of Dolz et al. [136], who proposed a suggestive annotation model for fully supervised infant brain MRI segmentation. By combining the outputs of 10 CNNs, they achieved state-of-the-art performance, showcasing how the weak models within an ensemble framework can correct each other's prediction results. However, to the best of our knowledge, due to the heavy computational load associated with such approaches, no previous studies have reported the application of ensemble learning and DCNNs to achieve semi-supervised learning in the field of image segmentation.

A generic semi-supervised image segmentation framework which integrates an ensemble technique is proposed. Firstly, the initial model is trained using annotated data and subsequently refined using unannotated data with pseudo masks. Different from Bai's method [100], the ensemble technique is used to reduce the negative influence of poor-quality pseudo masks. The proposed method was evaluated on a public skin lesion segmentation dataset. The results show that it outperforms both fully supervised learning method using only annotated data and Bai's method [100] by a large margin.

The rest of this chapter is organised as follows: Section 4.2 presents the methodology and the framework proposed in this work. Section 4.3 introduces the design of the experiments and the evaluation results of the proposed method on a public dataset. Finally, section 4.4 provides the discussion and conclusions of this chapter. Some additional works that related to the use of ensemble techniques are also introduced.

4.2 Methodology

In this section, the model architecture with an iterative training process is introduced. Then, the segmentation model structure and training approach to the initial model trained on all annotated images are given. This is followed by detailed information on the iterative training process, which demonstrates how the unannotated data improves the segmentation performance.

4.2.1 Model Architecture

An overview of the framework is illustrated in Fig. 4.1. The proposed framework is an iterative process, where the indices of iterations are represented by $\{L_0, \ldots, L_n\}$. A well-established encoder-decoder DCNN network (U-net [3]) is used as the basic segmentation method to train an initial model ($M0$) in L_0 using all annotated data, the detailed structure of the network is described in section 4.2.2.

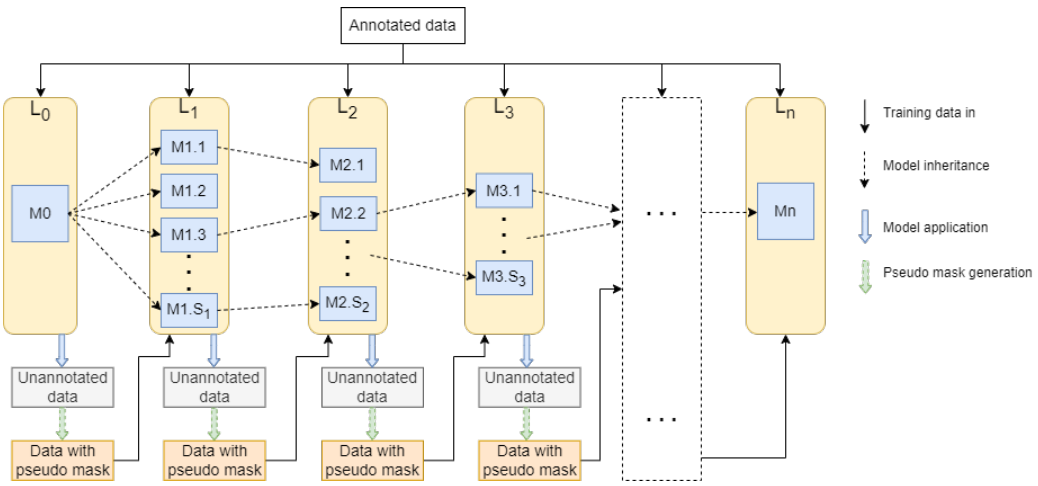

Figure 4.1: *Overview of the proposed framework.*

The model is then updated under an ensemble learning process. In detail, subsequent models from L_1 to L_n are trained based on both annotated data and a total number of N_0 unannotated data with pseudo masks generated from the previous iteration. The pseudo masks are generated using a weighted combination of outputs from all sub-models. The parameters of sub-models ($M1.1, \ldots, M1.S_1$) in L_1 are copied from $M0$. From L_1 onward, the number of

sub-models is reduced at each iteration and becomes one in the final iteration (L_n). Sub-models in the current iteration are randomly inherited from the sub-models of the previous iteration. The number of unannotated images ($\{N_0, \ldots, N_n\}$) used for training of each sub-model is gradually increased from level to level, following the rule of $N_n = \lfloor N_0/S_n + 0.5 \rfloor$ where $\lfloor \cdot \rfloor$ indicates the mathematical floor operation. Finally, M_n trained on all annotated data and unannotated data with pseudo masks is the final model used for segmenting any unseen data. More detailed descriptions of the ensemble strategy is presented in section 4.2.3.

4.2.2 Initial Supervised Segmentation Model

As described in chapter 2, U-net proposed by Ronnebergeret et al. [3] is a DCNN based encoder-decoder network which has achieved state-of-the-art performance for many image segmentation tasks in medical applications. Thus, it serves as the foundational segmentation network in this study.

In detail, the segmentation network in the proposed framework has an encoder and a decoder that both consist of several layers of feature maps by applying two 3×3 convolutional operations and one rectified linear unit (relu) [137] at each layer. In the encoding path, max-pooling with a stride of 2 is performed between two consecutive layers to achieve feature map down-sampling. Symmetrically, the decoder uses up-convolution to up-sample the feature map from the previous layer. Additionally, there are skip paths that concatenate the feature maps from the encoder to the corresponding layers of the decoder. A 1×1 convolution is used in the final decoder layer to convert the dimension of feature map to the number of classes. Subsequently, the softmax function is applied to map the activation values at each pixel position to the range of [0, 1].

The loss function used in this work combines both cross-entropy and Dice coefficient with equal weights, which outperforms the sole use of cross-entropy

as suggested by Liu et al. [138]. As a commonly used improvement, the residual block [83] was also added to the conventional block in the U-net for faster convergence. A validation dataset is used for determining the termination point of model training. Detailed parameter settings are provided in section 4.3.2.

4.2.3 Model Improvements Using Unannotated Data

Based on the initial segmentation model, the unannotated data is used to further improve the model. In detail, the model $M0$ is used to segment all N_0 unannotated data to generate segmentation outputs (probability of each pixel belonging to each class) which serve as pseudo masks. Then a random sub-set of these unannotated data with pseudo masks, together with the annotated data, are used to train a number of sub-models ($M1.1$,...,$M1.S_1$), which are called level one models in iteration L1. The initial parameters of these sub-models are the same and copied from $M0$. Then S_1 sub-models are generated initially, each with N_1 training data. The validation dataset for $M0$ is also used in subsequent levels to prevent the sub-models from model overfitting. For sub-model training from L_2 to L_n, the number of sub-models (S_n) is reduced as follows:

$$S_n = max(\lfloor \frac{S_1}{2^{n-1}} \rfloor, 1), \ for\ n > 1 \tag{4.1}$$

The whole framework stops training when S_n reaches 1. The training image contains a number of unannotated data and all annotated data. Hence the effect of the annotated data is gradually reduced when the number of unannotated images increases from level to level. This mechanism enables the model to gradually learn more information from unannotated data without a sudden performance drop.

From L_2 to L_n, the number of sub-models is reduced. They are randomly selected from the sub-models in the previous level. When all models in a level

have finished their training, a new set of pseudo masks (P_k, $k = 1,\ldots, N_0$) are generated for all the unannotated data based on a weighted combination of the output pixel-wise probability map ($M_{i,k}$) from all sub-models using:

$$P_k = \sum_{i=1}^{S_n} w_i M_{i,k} \tag{4.2}$$

The weight w_i for each sub-model is calculated as follows:

$$w_i = \sum_{j=1}^{R} B_{i,k}^{j} C_k^j \tag{4.3}$$

where $B_{i,k}^{j}$ is the binary map by applying a threshold of 0.5 to the pixel-wise probability map generated from the i^{th} sub-model for the k^{th} unannotated data. Superscript j is the index of pixels in an image that contains a total of R pixels. C_k^j is the summation of all pixel-wise probability maps generated from all sub-models for the k^{th} unannotated image. Intuitively, the weight of the i^{th} sub-model is calculated as the sum of probability values of the combined outputs from all sub-models within the image region predicted by the i^{th} sub-model. A larger weight indicates a greater agreement between the individual prediction of a sub-model and the combined prediction of all sub-models.

The weights of sub-models were scaled to the range of [0.1, 1] using Eq. (4.4), maximising the distance between the performance of the best and worst sub-models, and thus introducing effectively a 'relative reward' to reward the best sub-model the most, the worst sub-model the least, and apply a relative distribution between them.

$$w_i = \frac{w_i - min(w)}{max(w) - min(w)} \times 0.9 + 0.1 \tag{4.4}$$

Finally, weights are also normalised by dividing the sum of all weights. These new pseudo masks are used to train the sub models in the next level.

The task of sub-models is to learn features that are potentially different

from the initial supervised model. The outputs of these sub-models are then regulated and aggregated by the weighted combination process. These two processes work interactively to improve the segmentation performance. During this iterative process from level L_2 to L_n, the number of sub-models is decreased and the number of training images per sub-model is increased. The whole framework is terminated at level L_n when only one model is trained using all annotated and unannotated data.

4.3 Method Evaluation

4.3.1 Materials and Experiments

The proposed method was evaluated on "ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection" grand challenge dataset [81][82]. It provides 2594 images with manually annotated lesions (binary class). The image was pre-processed by using zero-mean normalisation (i.e. original intensity subtracted by the mean image intensity and divided by standard deviation). The original image size varies, which was resized to 128×128 to achieve a balance between computational efficiency and accuracy. The proposed method was also tested with larger image sizes (e.g. 256×256), which did not lead to significantly better performance but with much longer computational time, as the shapes of the lesions are not very complicated.

For the proposed method, the dataset was split into a training set and a test set with ratio of approximately 80%/ 20% (i.e 2094/ 500 images). Within the training set, the data was further split into a annotated set, an unannotated set and a validation set that contain 100, 1944 and 50 images respectively.

For comparison, a fully supervised model was trained using only the 100 annotated images (FS-100), and the full set of 2044 annotated images in the training set (FS-2044). The FS-100 result serves as a baseline to demonstrate the improvement of the proposed method by including unannotated data. The

FS-2044 result serves as the upper bound to indicate the best possible performance using all training images as annotated data. Bai's self-training method [100] as described in section 2.3.3 was also implemented for comparison.

4.3.2 Parameter Settings

For all methods, the basic network structures were the same, which was U-net with residual block (described in section 4.3.1). It consisted of 5 encoder and decoder layers respectively. The number of root features was 16 and doubled at the next layer in the encoder path and halved in the decoder path.

For the proposed method, the initial model was trained using 100 images for a maximum of 200 epochs with batch size of 10. The learning rate was 0.0001. To prevent over-fitting, a dropout operation with a 25% dropout rate was applied after each pooling layer. Due to the data size for sub-models being dynamic, the batch size used for sub-models was 1. For sub-model training, the maximum number of epochs was 50 and the learning rate was 0.0001. Ensemble learning was found to be capable of reducing bias and avoid over-fitting, therefore the dropout was not applied for faster convergence. The training process was stopped early when the loss value of the validation set was not decreased for 5 consecutive epochs. the number of sub-models in L_1 were set as 32, 16 and 8 with 6, 5 and 4 levels respectively. The results showed that a total number of 5 levels with the number of sub-models 16, 8, 4, 2, and 1 for L_1, L_2, L_3, L_4 and L_5 achieved the best performance in terms of balancing the computational time and segmentation accuracy.

Both fully-supervised methods FS-100 and FS-2044 were trained for 200 epochs with a learning rate of 0.0001. No early stopping was used, but a dropout operation with a 25% dropout rate was applied after every pooling layer. The batch sizes for FS-100 and FS-2044 were 10 and 28 respectively.

For Bai's method, the same 100 annotated images as in the proposed method were used to train an initial model and further refined it by includ-

ing 1944 unannotated images. Following the paper, the fully-connected CRF was applied on all pseudo masks in each iteration. After evaluating the CRF method on the validation set, the CRF parameters were set as $w_1 = 2$, $w_2 = 1$, $\sigma_\alpha = 2$, $\sigma_\beta = 3$, $\sigma_\gamma = 5$ (refer to [100] for more details). The self-training optimisation was performed for 3 iterations, with 50 epochs for each iteration. The proposed method was also experimented with more iterations, but it did not improve the performance further. Same as FS-2044, the batch size of this model was 28.

4.3.3 Results

This section reports quantitative results measured by Dice coefficient (DSC), Intersection over Union (IoU), accuracy, sensitivity, specificity and train time for performance comparison, as shown in table 4.1. These computations were conducted on a desktop PC equipped with an Intel Xeon W-2123 CPU operating at 3.6GHz, alongside a NVIDIA GTX 1070Ti GPU with 8GB of memory. The implementation of the code was carried out in Python, using the PyTorch deep learning framework.

Table 4.1: Comparison of the proposed proposed method to the FS-100 and FS-2044 models and Bai's method. Mean ± standard deviation values are reported.

Method	**DSC**	**IoU**	**Accuracy**	**Sensitivity**	**Specificity**	**Train time (s)**
FS-2044	0.872 ± 0.137	0.793 ± 0.171	0.952 ± 0.072	0.882 ± 0.156	0.974 ± 0.052	5926
FS-100	0.793 ± 0.190	0.693 ± 0.218	0.924 ± 0.103	0.859 ± 0.180	0.956 ± 0.076	2314
Bai's	0.817 ± 0.189	0.724 ± 0.215	0.933 ± 0.103	0.827 ± 0.203	0.970 ± 0.073	18822
Proposed	0.844 ± 0.156	0.755 ± 0.189	0.939 ± 0.091	0.887 ± 0.159	0.969 ± 0.062	27853

It is seen from table 4.1 that the FS-100 model and FS-2044 model produced the worst and best results respectively, as expected. Both Bai's method and the proposed method achieved better segmentation accuracy than the baseline FS-100 model, indicating successfully incorporating unannotated data for model improvement. More importantly, the proposed method is significantly ($p < 0.001$ based on paired t-test) better than Bai's method in terms of DSC, IoU and sensitivity measurements. For this dataset, the proposed method produced the best sensitivity values even better than the FS-2044

method but slightly lower specificity caused by more false positives.

The proposed method requires the training of a number of sub-models that leads to longer training time than Bai's method. Since the sub-models in earlier iterations (e.g. L1 and L2) only use a few images for training and the model parameters are inherited from previous models, the learning process converges rapidly. Hence the training time for the whole process is not tremendously high, particularly compared with a time-consuming manual annotation process.

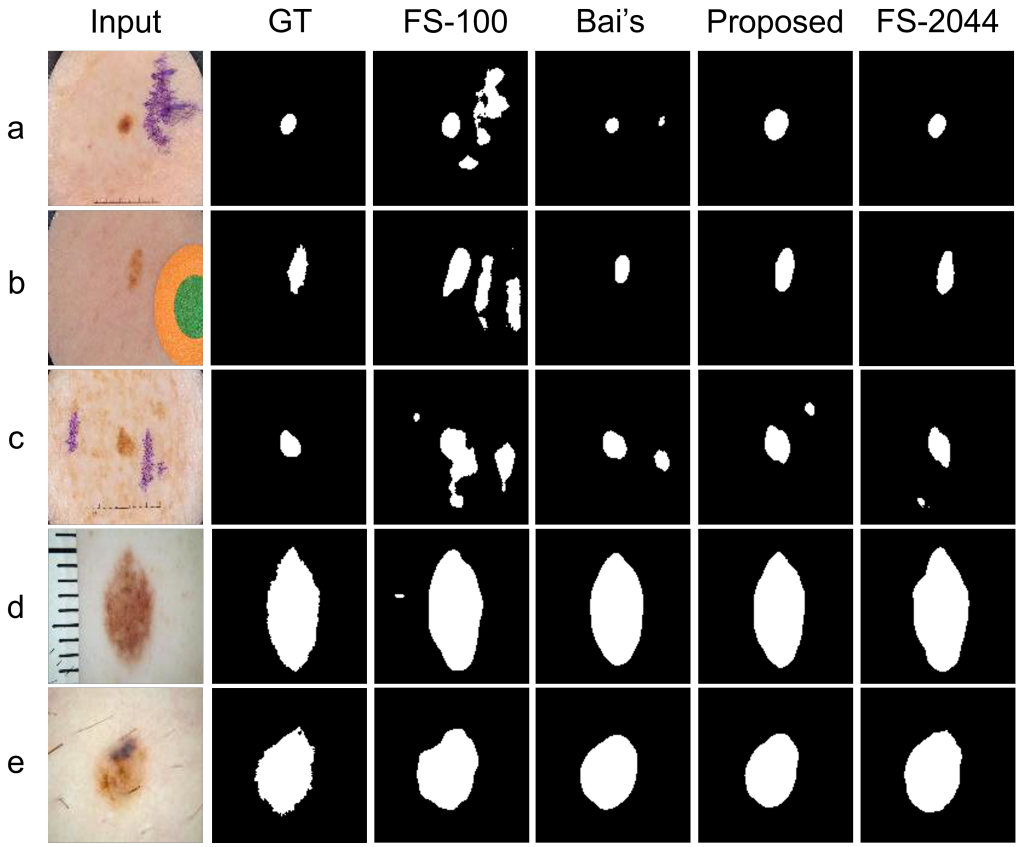

Figure 4.2: *Five examples of qualitative assessments for different models. The columns from left to right indicate the input image, ground truth, and four predictions generated by the FS-100 model, Bai's method, the proposed model, and the FS-2044 model, respectively.*

For qualitative assessment, some example images are shown in Fig. 4.2. Obviously, Fig. 4.2 (a)-(c) show that both Bai's method and the proposed method greatly improved the segmentation accuracy compared with FS-100 method by significantly reducing false positives for the images with high noise levels. This further proves the conclusion drawn from the quantitative results.

In Fig. 4.2 (d) and (e), it can be observed that there is minimal noise in the input images, and even FS-100 model, which is only trained on 100 annotated images, achieved relatively good segmentation results. This indicates that the initial model (FS-100) had already learned the general data distribution, but lacked noise resistance. By comparing it with Fig. 4.2 (a)-(c), this further confirms that adding unannotated data to the semi-supervised learning model allows it to extract richer information from the unannotated data, thus improving the noise resistance of the model. In addition, when compared with the ground truth, Bai's method excessively reduced the size of the target regions (more false negatives), where as the proposed method detected slightly larger region (more false positives).

4.4 Discussion and Conclusions

By integrating ensemble learning and DCNN, a generic semi-supervised learning framework is developed, which enables the improvement of fully supervised models by incorporating unannotated data. The framework was evaluated on a publicly available skin lesion dataset, and notable performance improvements were observed compared to a similar state-of-the-art semi-supervised learning method. The results demonstrated the superior effectiveness of the proposed method.

The capability to construct a fully automated medical image segmentation network for various types of medical image datasets has been achieved by leveraging the efficient interactive segmentation software developed in chapter 3 and the semi-supervised framework presented in this chapter. This approach only requires a large dataset and a limited number of annotated masks, which can be conveniently annotated by clinical experts using the provided interactive software.

In the next chapter, the focus is on further exploring the semi-supervised learning approaches for image segmentation. It has been discovered that many

segmented masks for medical images exhibit common geometric shapes. By employing image registration techniques, a segmentation mask can be generated by warping the mask of a template image onto the unannotated image space. This is achieved by leveraging the transformation information learned from both the template image and the unannotated image. Moreover, with the additional information provided by the image registration network, the pseudo-labelling-based semi-supervised image segmentation network can achieve improved performance with even fewer annotated images [21]. Consequently, the next chapter (chapter 5) introduces a comprehensive image registration approach capable of aligning datasets with significant deformations and utilising masks to guide the network's focus towards the mask region. The aim is to establish a joint image segmentation and registration framework as discussed in chapter 6.

Chapter 5

Mask Guided Image Registration

5.1 Introduction

This chapter introduces an alternative solution to the semi-supervised method proposed in chapter 4. The method proposed in this chapter adopts the idea of image registration. Despite decades of method development in image registration, some issues remain to be addressed. One of the key challenges is to cope with large image deformations.

Traditional image registration methods solve the problem of large deformation by multi-resolution strategy. Inspired by this strategy, several multi-resolution deep learning methods were proposed. Hering et al. [139] proposed a U-Net [3] based multi-resolution image registration framework. It uses multiple encoder-decoder networks to perform the estimation of displacement fields in different resolutions. The displacement field is a vector field that characterizes how each point in an image moves or changes position. These multi-scale networks are trained step by step from coarse to fine scales. The input of the higher resolution network is a down-sampled version of the target image and the warped and up-sampled source image produced by the previous lower resolution network. Mok et al. [109] proposed a Laplacian Pyramid based method

for large deformable image registration. Different from Hering's method, it uses several light encoder-decoder networks with skip connections of feature maps between them. Different networks focus on different image resolutions, the inputs of a higher resolution network are the source image, the target image and the up-sampled displacement field from the previous lower resolution network. The hyper parameters in these multi-resolution sub-networks need to be tuned and set differently to balance the weights of multiple loss function terms. Similarly, several U-Net based multi-resolution methods were proposed that combined feature maps from different resolutions [140] [141] [110]. Compared to the single-resolution deep learning methods and the traditional methods, all of these methods achieved significantly better results on images with large deformations.

Adopting the multi-resolution idea, in this chapter, a multi-resolution DCNN method for image registration of 2D and 3D medical images is proposed. Different from the aforementioned existing methods, the proposed method uses a single encoder to extract features from different image scales, and the decoder generates displacement fields for different resolutions. Then the displacement field in the finest scale is estimated by combining the up-sampled displacement fields from all coarser scales successively. The main contributions of this work compared to other DCNN methods are summarised as follows. (1) Without separate encoders for different resolutions, the proposed model uses a single encoder which is more compact and can be adjusted to different numbers of resolutions more flexibility. (2) The proposed method does not require a warped source image generated from the lower resolution as the input to the higher resolution layer. Instead, only the displacement field from the lower resolution is up-sampled and added to a learnable residual displacement field in the higher resolution, which requires shorter training time. (3) The residual displacement field design enables efficient and effective diffeomorphic deformation than the commonly used scaling and squaring method. (4) By incorporating a mask guided loss term, similar to [142] [113], the proposed method enables the

model to concentrate on the mask area, leading to improved local alignment.

The subsequent sections of this chapter are structured as follows: section 5.2 introduces the architecture of the proposed image registration network and the methodology used in this study. It also describes the training and the inference processes of the framework. Section 5.3 provides the details of the dataset and the experimental design, along with the evaluation results obtained from both 2D and 3D datasets. Section 5.4 provides the discussion and conclusions for this chapter. It leads to the method of integrating image registration and segmentation in the next chapter.

5.2 Methodology

5.2.1 Model Architecture

An overview of the proposed image registration framework is illustrated in Fig. 5.1. The network consists of an encoder and a decoder, and each has K levels (L_1,...,L_K) of scales. Fig. 5.1 is an example framework with 3 levels. In both the model training and inference processes, pairs of source and target images are merged into two-channel images, which are subsequently fed into the network. Each of the key components is introduced below.

Encoder: the encoder extracts multi-resolution features at different scales, each level of the encoder includes a residual block which consists of two 3×3 convolutional operators (3×3×3 for 3D) with stride of 1. Each of them is followed by a leaky rectified linear unit (Leaky ReLU) as the activation function [143]. The input of the residual block [83] is also added to the feature map learned after the second convolutional operator. A 1×1 (1×1×1 for 3D) convolutional operator with stride of 2 is applied to down-sample the feature maps between two consecutive levels.

Decoder: the decoder has the same number of levels as the encoder, and each level consists of one 3×3 (3×3×3 for 3D) convolutional operator

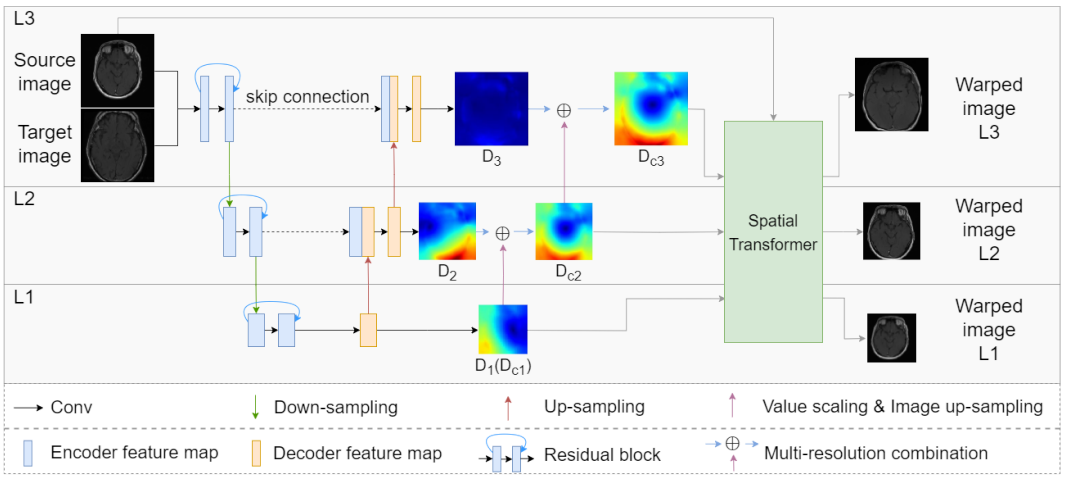

Figure 5.1: *Overview of the proposed multi-resolution image registration framework with 3 levels. Pairs of images (source and target) are input to the highest resolution level (level 3) to train the network. Each level estimates a displacement field (D_i), which is combined with the up-sampled displacement field from the lower level. The spatial transformer [9] warps the source image (S) to the target image (T) in each resolution to obtain a warped image ($f_D(S)$). The whole framework is updated by optimising the similarity between T and $f_D(S)$ with a smoothness term in each level.*

with stride of 1 and one 3×3 (3×3×3 for 3D) de-convolutional operator with stride of 2 as an up-sampling layer. Leaky ReLU is also used after both of them as the activation function. The connection between the encoder and the decoder in the lowest level (L_1) passes through a 3×3 (3×3×3 for 3D) convolutional operator followed by a Leaky ReLU. For all other higher levels, a skip connection is applied to concatenate the feature maps of the encoder to the decoder in the same level.

Residual Displacement Field: as illustrated in Fig. 5.1, a 3×3 (3×3×3 for 3D) convolutional operator is applied in each level of the decoder to directly estimate the displacement field (D_i) in the corresponding image resolution. During the model training process, the displacement field D_{c1} of the lowest resolution in level L_1 is estimated first. In level $L2$, the displacement field D_2 is calculated through the decoding process in that level. It is then added to the up-sampled version of D_{c1} to form the final displacement field D_{c2} for level L_2. To up-sample the displacement field (e.g. D_{c1}), the size of D_{c1} is firstly doubled using linear interpolation, followed by the scaling of the displacement values by a factor of 2. The higher levels follow the same procedure to successively

combine the displacement fields from the previous level. The displacement field D_i at each level can be considered as a residual displacement map, which shares a similar idea as the residual network. It enables a more efficient learning process and helps the diffeomorphic deformation as discussed in section 5.2.2.

Spatial Transformer and Image Warping: like most other image registration networks, the proposed method also employs a spatial transformer layer [9] to build an unsupervised image registration network. The spatial transformer layer consists of three main components. The first is the localisation network, which learns appropriate spatial transformations (i.e. displacement field) and is implemented using convolutional layers in this study. The second component is the grid generator, which defines the pixel-level mapping from the input feature map to the transformed feature map based on the estimated displacement field of the localisation network. The final component is the sampler, which performs the feature map transformation using the grid and employs linear interpolation.

For the implementation of unsupervised image registration, the model utilises randomly paired images (source image and target image). These images are input to the encoder and decoder of the network, generating feature maps. These feature maps are then passed through the localisation network, resulting in the computation of a displacement field. The displacement field is subsequently utilised by the grid generator and the sampler to facilitate the transformation of the source image. This transformation produces a warped image. The model then proceeds to update its weights iteratively through a learning process that optimises the similarity between the warped source image and the target image. The optimisation process can be expressed as follows:

$$\underset{D}{arg\,min}\,(L(f_D(S), T)) \tag{5.1}$$

where D indicates the displacement field. S and T are the source image and target image respectively. $f_D(S)$ indicates the warped image of warping S

using the displacement field D. L is the similarity measurement. Additional loss terms are added in our proposed method, which are described in sections 5.2.3 and 5.2.4.

5.2.2 Diffeomorphic Deformation

Diffeomorphic deformation is a critical property required in image registration. By estimating a smooth and invertible deformation field, it enables an accurate alignment of images while preserving the topology and spatial relationships between anatomical structures. This preservation of spatial relationships is crucial for obtaining reliable and meaningful registration results, so that image warping can be inverted using the inverse displacement field.

In some studies [144] [20] [109], the deformation field is treated as a static velocity field, and diffeomorphic deformation is obtained by adding a scaling and squaring layer along with multiple interpolations. The scaling and squaring layer [20] is used to calculate the displacement field from the flow field, providing a diffeomorphic registration. It refines the deformation field by iteratively scaling and squaring it t times, resulting in a smoother and invertible transformation. However, it requires several intermediate steps that is time consuming.

In the proposed method of this thesis, residual displacement fields are estimated at multi-level scales. Each of the residual displacement fields at different level only needs to perform a small deformation. By combining with the smoothness loss term (section 5.2.3), it helps achieving diffeomorphic registration without using the scaling and squaring technique. In the experiments (see section 5.3.3), it is shown that the proposed method is able to produce a similar diffeomorphic property to the scaling and squaring method in shorter training time.

5.2.3 Unsupervised Loss Functions

Given a source image S and a target image T, the objective of image registration is to transform S so that it can be aligned with T. The transformation in this case is represented by a displacement field D, which is used to warp the source image (denoted as $f_D(S)$). The image registration network is then optimised based on a similarity measurement between the target image and the warped source image. The global normalised cross correlation (GNCC), which is used to capture the global deformable information, is commonly used in medical image registration tasks (e.g. [145]). Local normalised cross correlation (LNCC) is also used in many studies, which focuses on local image similarities. Both methods are implemented for comparison, and their mathematical expressions are described below.

$$L_{sim} = -GNCC(x, y) \quad or \quad L_{sim} = -LNCC(x, y) \tag{5.2}$$

$$GNCC(x, y) = \frac{1}{N} \sum_{p \in \Omega} \frac{(x_p - \overline{x})(y_p - \overline{y})}{\sigma_x \sigma_y} \tag{5.3}$$

$$LNCC(x, y) = \frac{1}{N} \sum_{p \in \Omega} \frac{\sum_{i=1}^{w^2} (x_{p_i} - \overline{x_p})(y_{p_i} - \overline{y_p})^2}{(\sum_{i=1}^{w^2} (x_{p_i} - \overline{x_p}))(\sum_{i=1}^{w^2} (y_{p_i} - \overline{y_p}))} \tag{5.4}$$

where Ω indicates all the pixels (voxels) in the image. N is the total number of pixels in the images. x and y represent the warped image $f_D(S)$ and the target image T respectively. x_p and y_p are the intensity values at pixel p in x and y respectively. $\overline{x}$ ($\overline{y}$) and σ_x (σ_y) denote the mean and the standard deviation of the intensities of x and y. In Eq. (5.4), $\overline{x_p}$ and $\overline{y_p}$ are the mean image intensities of local regions around pixel (voxel) $p : \overline{x_p} = \frac{1}{w^2} \sum_{i=1}^{w^2} x_{p_i}$. p_i denotes the i^{th} pixel (voxel) of the w^2 (2D) or w^3 (3D) local region around p.

Another part of the loss function is a displacement field regularisation term. A smoothness regularisation loss [113] is calculated from each of the

(D_i), as expressed in Eq. (5.5):

$$L_{smooth}(D_i) = \|\bigtriangledown D_i\| \tag{5.5}$$

where $\bigtriangledown D_i$ denotes the approximate spatial gradients of displacement D_i using differences between neighbouring pixels (voxels).

The model learns an optimal displacement field by optimising Eq. (5.6).

$$arg\min_{D} \frac{1}{K}\sum_{i=1}^{K}(L_{sim}(f_{D_{ci}}(S), T) + \lambda_i L_{smooth}(D_i)) \tag{5.6}$$

where λ_i is the weight for the smoothness term on level i, which is used to balance the two terms. Both the similarity loss L_{sim} and the smoothness loss L_{smooth} are calculated on all K levels.

5.2.4 Mask Guided Loss Function

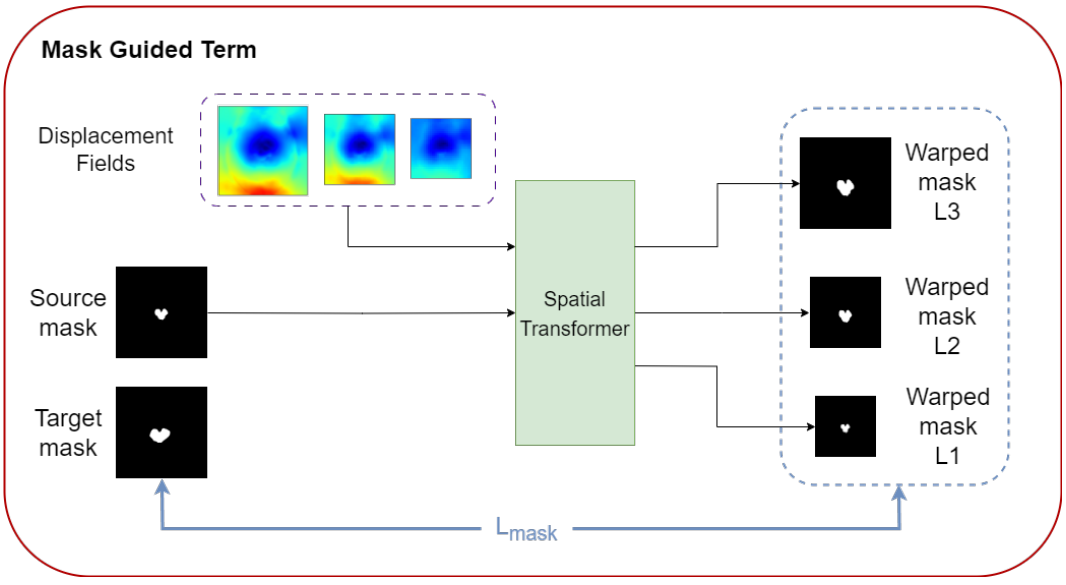

Figure 5.2: *The mask guided plug-in for the proposed registration model. The mask (mid-brain) of source image (brain MRI) is transformed into multiple resolutions using spatial transformer based on the displacement fields at different levels. The loss function comprises similarities between the mask of target image and the warped mask at each of the resolutions.*

In the context of medical image analysis, it is very common to focus on a small region of interest (ROI) (e.g. brain stem and cardiac atrium). In the task of

image registration, a globally aligned image pair does not guarantee a local optimal alignment. Hence, in the proposed method, the mask of ROI can be optionally incorporated into the network to guide the image registration to focus more on the ROI.

As illustrated in Fig. 5.2, a mask guided term is introduced to the multi-resolution framework during model training. In addition to the image similarity L_{sim} and smoothness loss L_{smooth}, a similarity loss between the warped mask of the source image and the mask of the target image in each resolution is applied. In each training iteration, the mask of the source image S_{mask} is transformed using the displacement fields at different resolutions (D_{c1}, D_{c2}, ..., D_{cK}) to produce warped masks ($f_{D_{c1}}(S_{mask})$, $f_{D_{c2}}(S_{mask})$, ..., $f_{D_{cK}}(S_{mask})$). The target mask is then down-sampled to match the corresponding image sizes, and the final loss is computed as the average of the similarities of all K scales. To address class imbalance, soft dice loss is used here instead of cross-entropy loss:

$$L_{mask}(x, y) = \frac{1}{K} \sum_{i=1}^{K} \frac{2\|x \cdot y\|}{\|x\|^2 + \|y\|^2} \tag{5.7}$$

where x, y represent the warped mask $f_D(S_{mask})$ and the target mask T_{mask}.

Finally, the optimisation term for model training incorporating the mask-driven loss is represented by Eq. (5.8). The network is then trained based on a training set that contains random pairs of source and target images, as well as their corresponding ROI masks as an optional setting.

$$arg \min_{D} \frac{1}{K} \sum_{i=1}^{K} (L_{sim}(f_{D_{ci}}(S), T) + L_{mask}(f_{D_{ci}}(S_{mask}), T_{mask}) + \lambda L_{smooth}(D_i)) \tag{5.8}$$

5.2.5 Model Inference

Once the model is trained, it can be applied to estimate the displacement fields between new pairs of source and target images. As depicted in Fig.

5.3, by providing a pair of images, namely a source image S and a target image T, the model can generate predictions for the displacement field D, which represents a mapping that transforms the source image S to the target image T. Consequently, the model can utilise the displacement field to warp the source image, resulting in a warped source image denoted as $f_D(S)$ using the spatial transformer. Note that, different from the training process, the mask of the target image is not required in the model inference process. If the source image includes an annotated mask S_{mask}, the model can also apply the mapping $f_D(S_{mask})$ to warp it into the target space. The warped mask of the source image can be considered as the segmentation mask of the target image, assuming the image registration is sufficiently accurate. The presence of the mask does not affect the registration result. The network automatically focuses more on the ROI as learned from the training process. It is demonstrated by the results presented in the next section.

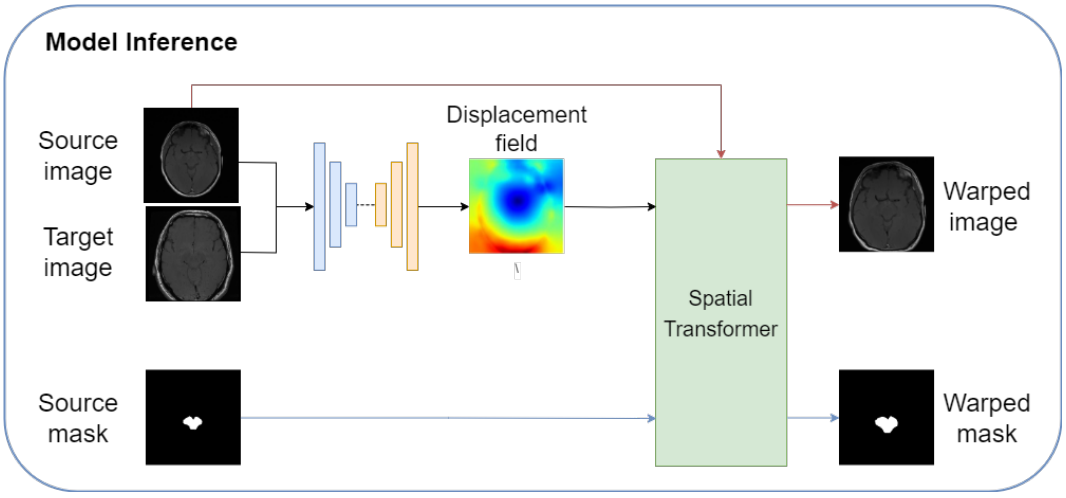

Figure 5.3: *The model inference process of the proposed framework. The registration model generates a displacement field from a source-target image pair. The spatial transformer warps the source image using this field, producing a warped image (red route). If a segmentation mask exists, it is also warped using the same displacement field, generating a predicted segmentation outcome for the target image (blue route).*

5.3 Method Evaluation

5.3.1 Datasets

The proposed method was evaluated on a widely used public 3D brain MRI dataset and a more challenging 2D brain MRI dataset collected locally.

The public 3D brain MRI dataset contains 414 T1-weighted brain scans with 4 manually annotated anatomical regions (i.e. cortex, subcortical gray matter (SGM), white matter (WM) and cerebrospinal fluid (CSF)) is from the OASIS dataset [146]. The images were pre-processed by the authors of HyperMorph [147] using FreeSurfer [148], which include spatial normalisation, bias-correction and skull-stripping. The size of the images after pre-processing is $160 \times 192 \times 224$, and then were further resized to $96 \times 96 \times 96$ to reduce the memory consumption and training time. The same pre-processed images were used consistently for a fair comparison of all methods.

The local 2D brain MRI dataset contains 820 T1-weighted slices with manually annotated mid-brain as the target of interest. These slices are from slightly different cross-sectional brain regions of different subjects, and were acquired by different MRI scanners. Except for resizing the image to 256×256, no other pre-processing was performed to these images. Very large spatial differences across different subjects were observed in this dataset which is more challenging than the 3D brain dataset.

5.3.2 Experimental Methods

The proposed method was compared to one of the most widely used traditional image registration method Demons [149, 150] and a state-of-the-art deep learning method VoxelMorph [113]. The official implementation of Demons in SimpleITK toolkit [151] was used. For VoxelMorph, the published code by the authors was used with a slight modification to ensure the lowest image resolution is comparable among all methods for a fair comparison.

For the 3D dataset, the data was divided into training, validation and test sets with the ratio of approximately 60%, 5% and 35% (i.e. 244, 20 and 150) respectively. For the 2D brain dataset, the images were divided into training, validation and test sets with the ratio of approximately 75%, 5% and 20% (i.e. 600, 40 and 180) respectively.

In the training process, each of the training images was selected as the target image in turn, and the source image was randomly selected. This approach enabled randomised pair selection, as well as ensured that the model learned from a diverse range of image combinations throughout the training process.

To evaluate all methods, 5 randomly selected images from the test set were used as the source images, and the remaining test images as the target images to evaluate the methods. Therefore, there are 725 paired 3D images and 875 paired 2D images for testing. This ensures all methods were applied to the same set of paired images for a fair comparison.

Two models MrRegNet-G and MrRegNet-L based on the proposed network, which were trained using GNCC (global) and LNCC (local) as the similarity loss respectively, were tested and compared to Demons and VoxelMorph. The hyper-parameters were tuned using the validation dataset. The 2D dataset and 3D dataset had different hyper-parameters as detailed below.

Parameter Settings for 2D Brain Dataset

For both MrRegNet-G and MrRegNet-L, the learning rate was set to 0.001, and the training epochs was 200. The batchsize was set to 10. The number of levels was determined based on the image size, resulting in 5 levels. In the case of MrRegNet-G, the smoothness term weight (λ_i) varies dynamically across levels, ranging from 128 to 8 and halved at each level ($\lambda_1 = 128$, $\lambda_2 = 64$, $\lambda_3 = 32$, $\lambda_4 = 16$, $\lambda_5 = 8$). This approach assigned higher weights to lower resolution levels to enforce higher levels of smoothness in coarse levels. For MrRegNet-L, the smoothness term weights are the same for all levels, as the

LNCC prioritises local regions rather than global information, ensuring equal smoothing weight for each local region. The lowest resolution of the 2D image is calculated as $W/2^{N-1} \times H/2^{N-1} = 256/2^{5-1} \times 256/2^{5-1} = 16 \times 16$, where W and H represent the width and height of the 2D image, respectively. In this case, N was set to 5. Consequently, the local region size w for LNCC was set to 9 to encompass sufficient information for the lowest resolution, and the smoothness term weight λ was set to 10. In order to ensure a fair comparison, the Demons method also adopted the pyramid structure with a 5-level configuration on the 2D dataset, while the other settings remain at the default values.

The training parameters were tuned for VoxelMorph-G and VoxelMorph-L using GNCC and LNCC as the similarity loss separately. Similar to MrRegNet, the learning rate was set to 0.001, and the training was conducted for 200 epochs. The batchsize was also set to 10. After parameter tuning using the validation set, the weight of the smoothness term was set to 10 for VoxelMorph-G and 1 for VoxelMorph-L. Specifically, for VoxelMorph-L, the local region size w for LNCC was set to 9 to ensure a fair comparison to MrRegNet-L.

Furthermore, for the proposed MrRegNet, additional experiments were conducted to compare the performance of MrRegNet-G-SS and MrRegNet-L-SS, which include the addition of scaling and squaring layers. All hyper-parameters remained the same as those without the scaling and squaring layers (diffeomorphic deformation). In order to achieve diffeomorphic image registration, the parameter t was set to 5, indicating that the displacement field would undergo scaling and squaring operations 5 times.

Parameter Settings for 3D Brain Dataset

Only the results of MrRegNet-G was reported on this dataset, as it was observed that MrRegNet-G on 2D dataset produced more reliable image alignment results than MrRegNet-L. Additionally, several experiments on multi-classes are conducted to discover the usage of mask guided loss. In these

experiments, a learning rate of 0.0001 was set to facilitate smoother training. The training epochs was 200. The batchsize was set to 1. To ensure that the lowest resolution was not too small for registration displacement field learning, a 4-level network structure was employed on an image with size of $96 \times 96 \times 96$. This configuration allowed the lowest resolution of $12 \times 12 \times 12$ at the deepest level. Like the experiments conducted on the 2D dataset, the weights assigned to the smoothness term were dynamic. However, in this case, the weights ranged from 16 to 2, with $\lambda_1 = 16$, $\lambda_2 = 8$, $\lambda_3 = 4$, and $\lambda_4 = 2$. Same as the 2D dataset, the Demons employed 4 layers for a fair comparison, the remaining settings were set as default.

The VoxelMorph-G model was also assessed on this dataset, which used GNCC as the similarity loss function. The learning rate was set to 0.0001, the batchsize was set to 1, and the training lasted for 200 epochs, same as the settings of MrRegNet-G. The weight of the smoothness term was set to 1.

Experiments for Mask Guided Loss and Diffeomorphic Deformation

Besides testing the performance of the proposed MrRegNet model, additional experiments were conducted to evaluate the advantages provided by the mask guided loss term. The method is named as MrRegNet(mask). As mentioned in section 5.3.1, the 2D brain data contains annotations for only one class in the central region of the brain, while the 3D brain data has annotations of four classes that are distributed around the whole brain. These two datasets help to test the proposed method in different scenarios (i.e. binary mask, multi-class mask, small mask and large mask).

Furthermore, the widely used scaling and squaring method was added to the proposed model for the purpose of preserving diffeomorphic deformation. The method is named as MrRegNet-SS. As stated in section 5.2.2, the proposed network learns a residual displacement field in each scale, which helps to retain the diffeomorphic property. Hence, this experiment helps to demonstrate the proposed method can achieve similar diffeomorphic deformation without using

the time consuming scaling and squaring method.

In combination with the local (-L) and global (-G) similarity measurements, different variants of the proposed network were implemented for comparison, including MrRegNet-G, MrRegNet-L, MrRegNet-G(mask), MrRegNet-L(mask), MrRegNet-G-SS, MrRegNet-L-SS, MrRegNet-G-SS(mask) and MrRegNet-L-SS(mask). These were compared to VoxelMorph and Demons methods.

Evaluation Metrics

The commonly used evaluation metric, global normalized cross-correlation (NCC), was used to measure the quality of image registration. It measures the similarity between the warped source image and the target image. The value range is between 0 to 1. A higher value indicates greater similarity between the two images.

Besides, Dice coefficient (DSC) on the annotated anatomical regions between the warped image and the source image was used to measure the quality of registration in the masked area [109] [110]. DSC ranges from 0 to 1, where 1 indicates perfect agreement between two masks. The DSC of annotated regions between the source image and the target image before registration was reported as a baseline.

In addition, the rate of non-positive value of Jacobian determinant $\|J_D\| \leq 0$ and standard deviation of Jacobian determinant on the estimated displacement field $s(\|J_D\|)$ were also used to measure the quality of the diffeomorphic property. A lower number of non-positive value in Jacobian determinant indicates a better diffeomorphic property. Smaller standard deviation of Jacobian determinant indicates the displacement field is smoother and locally invertible.

5.3.3 Results

2D Brain MRI Data

The quantitative results on the 2D brain dataset are reported in table 5.1. First of all, without the mask guided term and the scaling and squaring layer, the proposed methods (MrRegNet-G and MrRegNet-L) outperform all other methods in terms of GNCC, DSC and $s(\|J_D\|)$. Furthermore, when comparing local and global similarity loss terms within the same method, it can be seen that the local similarity loss (-L) achieved a better DSC than the global similarity loss (-G) for both VoxelMorph and MrRegNet. This suggests that without the guidance of a segmentation mask, local similarity performed better in local alignment. However, the GNCC values of the -L methods are lower than the -G methods for both VoxelMorph and MrRegNet, which indicate poor global alignments using the local similarity loss.

Table 5.1: Quantitative evaluation on 2D dataset. The baseline results are the measurements before image registration. The results of Demons, VoxelMorph and MrRegNet are compared. "-G" and "-L" represent the loss function GNCC and LNCC respectively. "-SS" indicates the method using scaling and squaring method. "(mask)" indicates a mask guided loss term was added to the model. The mean ± standard deviation values of global normalized cross-correlation (GNCC), Dice coefficient (DSC), ratio percentage non-positive value $\|J_D\| \leq 0$ and the standard deviation of Jacobian determinant $s(\|J_D\|)$ are reported for each method. The reported values are presented as the mean ± standard deviation.

Method	**GNCC**	**DSC**	$\|\boldsymbol{J_D}\| \leq \mathbf{0}$	$\boldsymbol{s(\|JD\|)}$
Baseline	0.39±0.05	0.66±0.15	n/a	n/a
Demons	0.75±0.09	0.68±0.17	0.00±0.00	0.22±0.07
VoxelMorph-G	0.77±0.03	0.71±0.14	0.85±0.38	0.37±0.03
VoxelMorph-L	0.60±0.09	0.72±0.16	0.51±0.10	0.33±0.04
MrRegNet-G	0.80±0.03	0.77±0.10	0.01±0.03	0.14±0.02
MrRegNet-L	0.78±0.04	0.80±0.10	0.01±0.03	0.18±0.02
VoxelMorph-G (mask)	0.65±0.04	0.82±0.07	0.05±0.07	0.19±0.01
VoxelMorph-L (mask)	0.50±0.06	0.87±0.07	0.23±0.08	0.22±0.02
MrRegNet-G (mask)	0.76±0.04	0.86±0.06	0.07±0.07	0.17±0.02
MrRegNet-L (mask)	0.73±0.05	0.87±0.08	0.23±0.13	0.21±0.02
MrRegNet-G-SS	0.80±0.03	0.78±0.10	0.01±0.01	0.15±0.03
MrRegNet-L-SS	0.77±0.04	0.79±0.09	0.02±0.04	0.19±0.03
MrRegNet-G-SS (mask)	0.78±0.04	0.86±0.06	0.05±0.07	0.16±0.03
MrRegNet-L-SS (mask)	0.75±0.04	0.85±0.09	0.21±0.14	0.22±0.03

Based on the results of the masked versions of both VoxelMorph and MrRegNet methods in table 5.1, it is seen from the DSC values that the mask-guidance loss is capable of improving the local alignment in the masked region. Moreover, when examining the GNCC values, it can be observed that the mask-guided loss results in reduction of the global registration performance for all methods. In the case of VoxelMorph, the performance of VoxelMorph-G (mask) and VoxelMorph-L (mask) show a noticeable decrease (10-13%) from 0.78 to 0.65 and from 0.60 to 0.50 respectively. In contrast, the proposed methods MrRegNet-L (mask) and MrRegNet-G (mask) achieved a significantly higher performance in local alignment (5-10% higher DSC values) with a relatively small decrease in global alignment (3-4% lower GNCC values). The MrRegNet-G (mask) method is preferred in terms of balancing between the global and local alignments, resulting in GNCC of 0.76 and DSC of 0.86.

Regarding diffeomorphic properties, as evidenced by the absence of non-positive values and lower standard deviation of the Jacobian determinant showing in table 5.1, without the mask guided term, the proposed method exhibited superior performance compared to VoxelMorph in maintaining a good diffeomorphic property of the displacement field. After adding the mask guided term, VoxelMorph (mask) achieved similar performance as the proposed MrRegNet (mask). However, it can be observed later from Fig.5.5 that this improvement was a result of the unsuccessful global alignment of the images, leading to a reduction in displacement values. For the proposed MrRegNet, after the inclusion of the mask guided term, when observing the standard deviation of Jacobian determinant in table 5.1, it can be concluded that the smoothness did not noticeably worsen. However, the non-positive value of Jacobian determinant indicates that due to the increased emphasis on local region alignment, slightly more folding pixels occurred than without using the mask-guided term.

Furthermore, table 5.1 also provides the results of the proposed methods by adding the scaling and squaring layer with $t = 5$ (-SS). The results indicate

that incorporating this layer did not significantly improve the performance of the model in terms of registration accuracy, and there was no notable improvement in the diffeomorphic property. These findings suggest that the proposed method can achieve diffeomorphic registration effectively without using the scaling and squaring method.

Follow on the above quantitative analyses, some qualitative analyses of the results are conducted by visualising some examples.

Fig. 5.4 shows three registration examples by applying Demons, VoxelMorph and MrRegNet without the mask guided term. Specifically comparing row (a) and row (b) of Fig. 5.4, it is evident that while Demons is capable of handling registration involving significant deformations, it fails to achieve accurate alignment along the boundary (row (a)) and in detailed regions (top area of the brain in row (b), indicated by the red arrow). VoxelMorph-G demonstrates the ability to align images with large deformations, but it results in a significant pixel folding issue (row (a), top region highlighted by the red arrow). Moreover, VoxelMorph-L fails to align the source and target images correctly. In row (c), all the methods produce similar results to register the images with small deformations. By employing a local similarity loss (-L), both VoxelMorph and the MrRegNet can focus more on the local region (as indicated by the blue arrows in row (c)). Overall, Fig. 5.4 demonstrates that the proposed MrRegNet method is able to achieve consistent performance for both large and small deformations, which is preferable than the other compared methods.

Regarding the mask guided term, Fig. 5.5 and Fig. 5.6 show the visualisation outcomes of an example paired images by applying VoxelMorph and MrRegNet respectively. By observing the masks and the heatmaps and grid views of the displacement field, they demonstrate that for both methods, the inclusion of the mask-guided term results in capturing fine displacements within the masked mid-brain region, and the generated masks have a more similar shape to the target mask. This signifies that the model focuses more

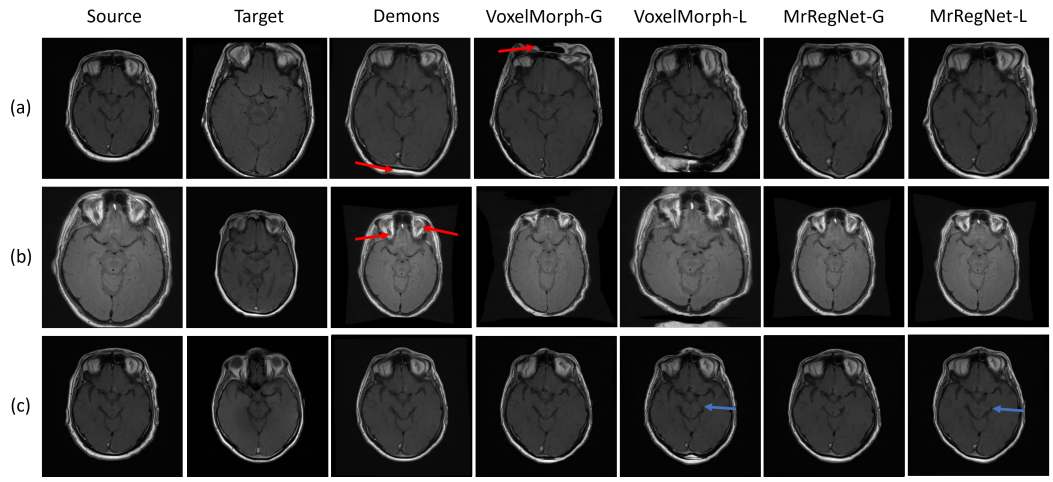

Figure 5.4: *Visualisation results of different registration methods without mask guided loss term on the 2D brain dataset. All methods, except the VoxelMorph-L (a) and (b), achieved good global alignments. The red arrows in rows (a) and (b) indicate specific small regions where Demons and VoxelMorph-G exhibited poor alignment. In row (c), the blue arrows point to the mid-brain region of VoxelMorph-L and MrRegNet-L, showcasing effective local region alignment. Further details can be found in Section 5.3.3.*

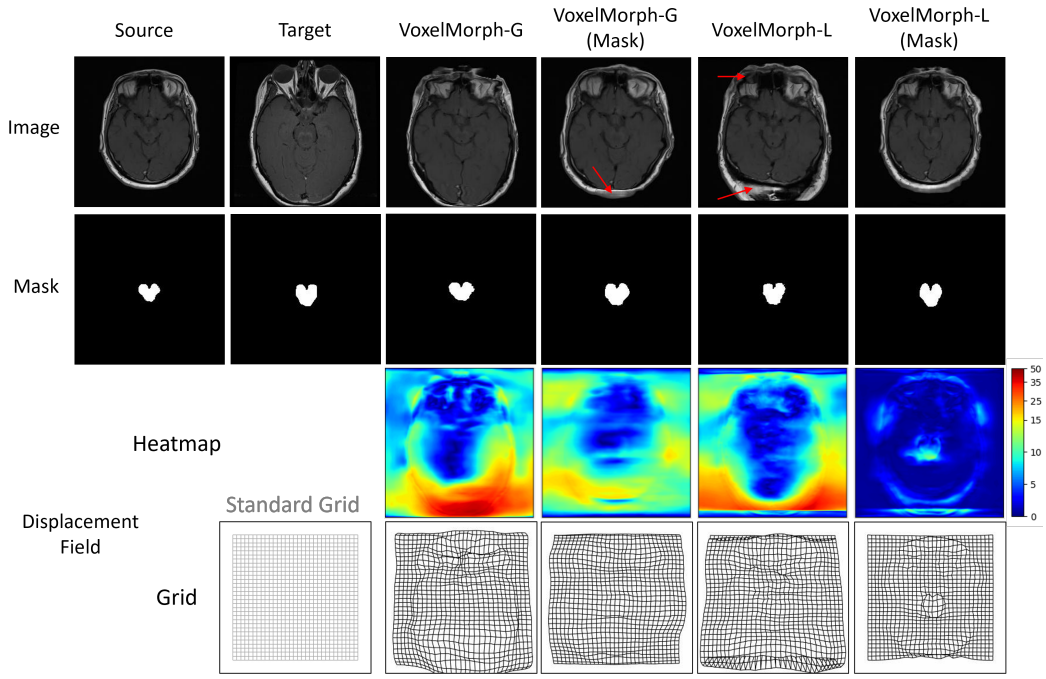

Figure 5.5: *The visualisation showcases registration examples of the methods based on VoxelMorph on a 2D image pair. The first and second rows depict the images and masks, respectively. The third row presents a heatmap of the estimated displacement field for each method (the higher the value, the larger the pixel shifts). The fourth row displays the deformation grids, including the grid before registration, and the grids of the displacement fields after registration.*

on aligning the masked region.

For the VoxelMorph method presented in Fig. 5.5, both VoxelMorph-L

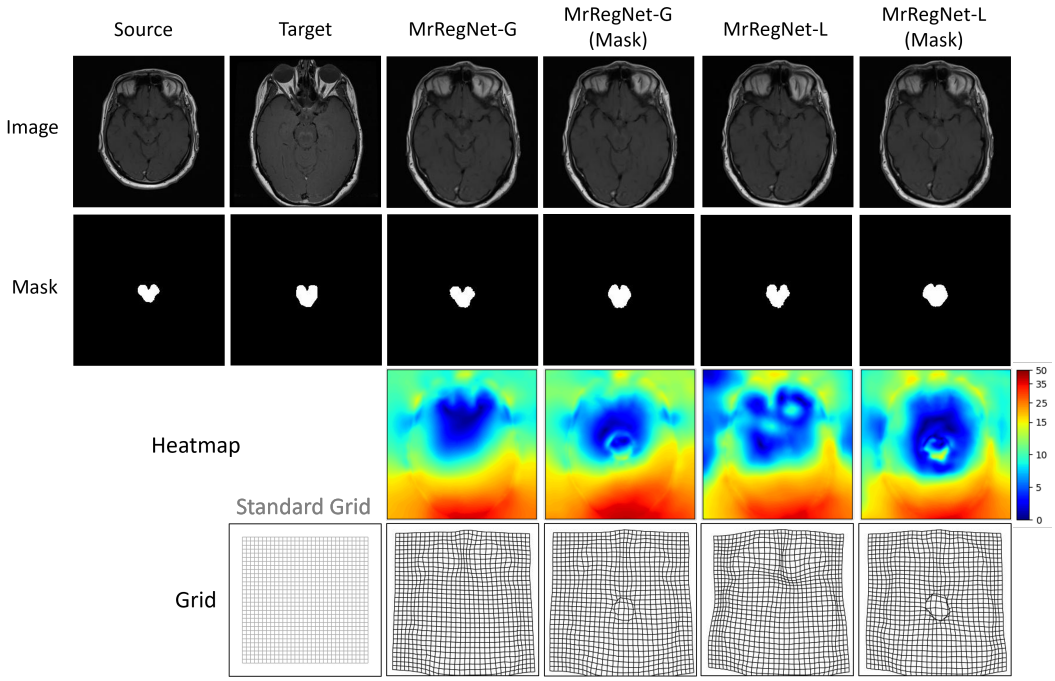

Figure 5.6: *Visualisation of the same example image as shown Fig. 5.5 based on the proposed MrRegNet. The layout is the same as in Fig. 5.5.*

and VoxelMorph-L (mask) show difficulties in handling largely deformed images. Although VoxelMorph-L (mask) achieved an acceptable alignment in the mid-brain region due to the mask guided term, its focus remained primarily on the masked region, resulting in a mis-alignment of the whole brain region. Furthermore, the heatmaps and grid views exhibit notable discrepancies depending on the chosen loss function, indicating VoxelMorph's sensitivity to these settings. In comparison to the proposed method in Fig. 5.6, the displacement fields are also less smoother.

The visualisation of the same example using the proposed method is shown in Fig. 5.6. Better than VoxelMorph, all the MrRegNet variants achieved acceptable results. The displacement fields of these methods are similar to each other and smooth, despite their different loss functions. This indicates better robustness and stability of the proposed method than VoxelMorph. The heatmaps and grid views of MrRegNet-G (mask) and MrRegNet-L (mask) contain some fine movements in the mid-brain area, indicating the significant impact of the mask-guided term in aligning local regions. Additionally, it is observed that MrRegNet-L tends to focus more on the local region compared

to MrRegNet-G, as expected. However, with the inclusion of the mask-guided term, MrRegNet-G (mask) is able to prioritise the masked region without significantly affecting other areas. However, MrRegNet-L (mask) loses some focus on the region surrounding the masked region, resulting in poorer registration performance compared to MrRegNet-G (mask). This finding is consistent with the conclusion of the quantitative results in table 5.1.

In summary, based on both the quantitative and qualitative analyses, the proposed method MrRegNet-G (mask) achieved the best performance in terms of balancing on global alignment, local alignment and diffeomorphic properties. It also leads to higher quality alignment than other methods (i.e. Demons and VoxelMorph) on visual inspections.

3D Brain MRI Data

Table 5.2: Quantitative evaluation on 3D dataset. The Baseline, Demons, VoxelMorph-G, and MrRegNet-G methods remain consistent with those listed in table 5.1. The addition of "(masks)" indicates the mask guided loss was calculated on all classes during model training. The Global Normalized Cross-correlation (GNCC), Dice coefficient (DSC) for each class and the average score, percentage non-positive value $\|J_D\| \leq 0$ and standard deviation of Jacobian determinant $s(\|J_D\|)$ are reported for each method.

Method	**GNCC**	**DSC**					$\|J_D\| \leq 0$	$s(\|JD\|)$
		Mean	**Cortex**	**SGM**	**WM**	**CSF**		
Baseline	0.67±0.14	0.25±0.11	0.26±0.08	0.26±0.16	0.35±0.12	0.11±0.13	n/a	n/a
Demons	0.95±0.02	0.62±0.06	0.57±0.05	0.46±0.07	0.72±0.05	0.7510±0.10	0.01±0.01	0.13±0.02
VoxelMorph-G	0.93±0.02	0.55±0.06	0.47±0.04	0.61±0.09	0.64±0.04	0.49±0.15	0.16±0.19	0.20±0.04
MrRegNet-G	0.94±0.02	0.62±0.06	0.51±0.04	0.67±0.08	0.67±0.04	0.62±0.14	0.13±0.09	0.21±0.02
MrRegNet-G (mask)	0.94±0.02	0.67±0.05	0.53±0.04	0.74±0.08	0.70±0.04	0.72 ±0.11	0.43±0.15	0.27±0.02

The quantitative results on the 3D brain dataset with multi-class masks are shown in table 5.2. As concluded from the experiments on 2D dataset, the global similarity loss works better with the mask guided loss. Hence, only Demons, VoxelMorph-G, MrRegNet-G and MrRegNet-G (mask) are compared for the 3D dataset. By examining the GNCC and mean DSC values, all the methods achieved good registration results on this dataset. This is due to that the 3D dataset contains relatively small image deformations compared to the 2D dataset.

Among the compared methods without mask guidance, Demons achieved

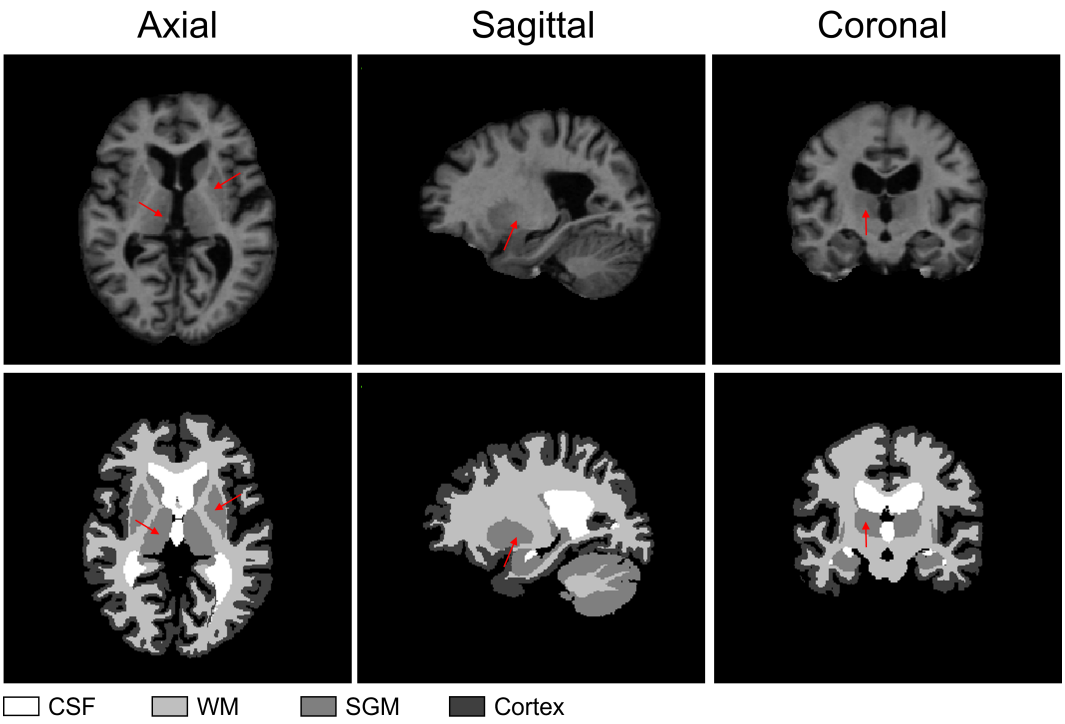

Figure 5.7: *3D mask template showcasing images and masks in axial, sagittal, and coronal views. Arrows indicate the subcortical gray matter (SGM) region in both the image and the mask to highlight the intensity variations in the SGM region.*

overall the best performance on GNCC, mean DSC and Jacobian determinant measurements. By looking at the DSC score for each masked brain region in table 5.2, Demons achieved better results on Cortex, WM and CSF region, but a worse result on SGM. It is hypothesised that the lower performance on SGM region is due to the inhomogeneous intensities within the region as shown in Fig. 5.7, which makes it challenging for non-learning based method like Demons. In contrast, the deep learning based methods (voxelmorph-G and MrRegNet-G) are more robust in feature learning to cope with intensity variations. When comparing VoxelMorph-G and the proposed MrRegNet-G, it is obvious that MrRegNet-G performed better on all metrics, except a similar performance on $s(\|JD\|)$.

More importantly, when incorporating the mask guided term into the proposed method (MrRegNet-G (mask)), it is seen from table 5.2 that while the overall GNCC measure is similar to Demons (0.94 vs. 0.95), but the mean DSC value (0.67) is significantly better than all the other methods ($\leq$0.62).

The smoothness of the displacement field is slightly affected by adding the mask guidance ($s(\|JD\|)$: from 0.20 to 0.27). The number of folding voxels were also increased ($\|J_D\| \leq 0$: from 0.13 to 0.43), due to the increased movement on local regions. Unlike the tests on the 2D dataset, the addition of the mask guided term in this case does not lead to a drop in GNCC measure, indicating that the model possesses the capability to handle both global and local deformations when the global deformation is not excessively large.

Visualisation of Displacement Fields

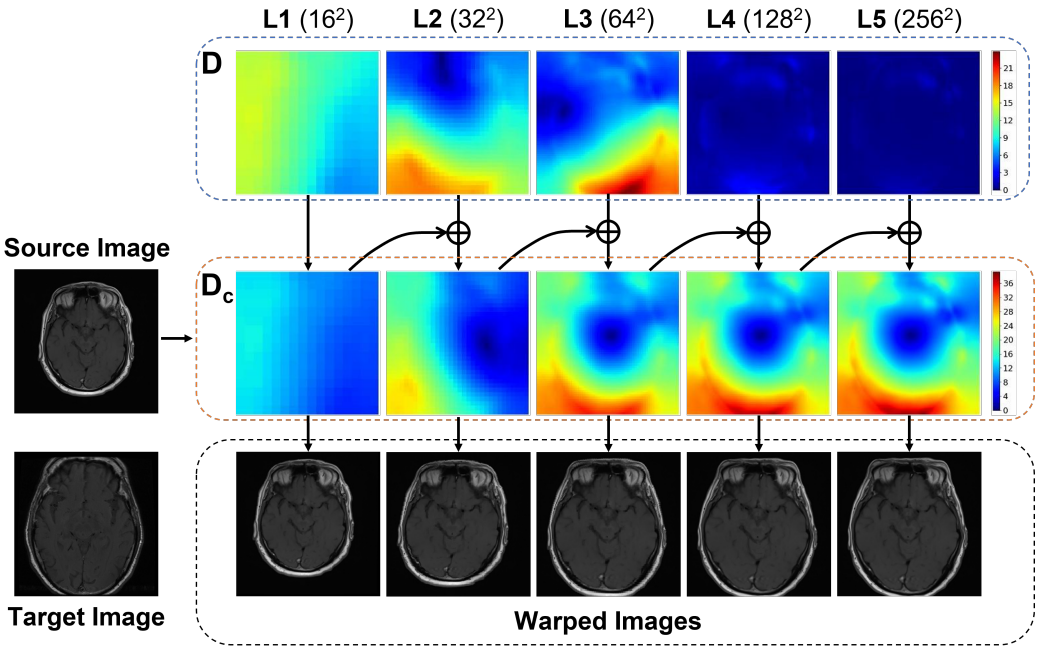

Figure 5.8: *An example of registering a source image to a target image using MrRegNet-G. The heatmaps show the scaled residual displacement fields (D) and the scaled combined displacement fields (D_c) at different levels ($L1$, $L2$, ..., $L5$), with resolutions of 16^2, 32^2, 64^2, 128^2, and 256^2. The values of displacement fields are resized to 256^2 and scaled by 2^{5-K}, where K represents the level number. The warped images are generated by applying the scaled D_c to the source image. The colour bar indicates the magnitude of pixel shift, with higher values corresponding to larger shifts. Note that each row of the colour bar has the same values, indicating consistent pixel shift magnitudes across different scales.*

While the $s(\|JD\|)$ and $\|J_D\| \leq 0$ measures can quantify the smoothness and folding pixels (voxels) of the displacement field, it is also desirable to visually inspect the displacement fields. Fig. 5.8 shows an example of the image warping process using the displacement fields obtained from different

resolutions. By examining the heatmap of the residual displacement field, it is evident that different resolution levels capture different scales of movements. Notably, the first three resolutions primarily control large-scale deformations in different regions, while higher resolutions handle more intricate details, such as the boundary of the brain.

By combining these resolution-specific displacement fields, the resulting combined displacement field encompasses deformations across a wide range of scales, from coarse to fine. This comprehensive approach ensures that all types of deformations are accounted for in the final result. Furthermore, the process of combining the displacement fields effectively smooth out the overall displacement field while preserving its diffeomorphic property.

Computational Time

The computational time is also a key factor when comparing different methods, hence reported in table 5.3 for both 2D and 3D datasets by applying all compared methods. The computations were performed using a GPU server with an Intel E5-2620 v4 CPU running at 2.10GHz and a NVIDIA GTX 1080Ti GPU with 11GB memory. The code was implemented in Python using the PyTorch deep learning framework.

The table shows that despite a long training time, deep learning-based methods (VoxelMorph and MrRegNet) have significantly faster inference speeds compared to the traditional Demons algorithm. Even when executed on a CPU, these methods surpass Demons by more than 15 times in speed on 3D data. The acceleration is even more remarkable, exceeding 40 times, for the 2D data. By leveraging GPU acceleration, both VoxelMorph and MrRegNet achieve inference time of below one second (less than 1/10 of a second for 2D images).

On both 2D and 3D datasets, the proposed method MrRegNet is approximately 30% slower in training compared to VoxelMorph. This is attributed to the utilisation of multi-scale registration and the inclusion of multiple spatial

Table 5.3: The computational times for different methods on both 2D and 3D were provided, which includes model training time (GPU), total (per epoch), and inference time (CPU and GPU). The time is measured in seconds.

Dataset	Method	Training Time (s)	Inference Time (s)	
			CPU	GPU
2D	Demons	n/a	8.39	n/a
	VoxelMorph	1309 (6.55)	0.07	0.03
	MrRegNet	1713 (8.57)	0.16	0.03
	MrRegNet-SS	2288 (11.44)	0.20	0.04
3D	Demons	n/a	36.76	n/a
	VoxelMorph	27450 (137.25)	1.50	0.34
	MrRegNet	36291 (181.46)	2.32	0.45

transformer layers. Therefore, the inference speed of the proposed method is also slightly slower than VoxelMorph. Furthermore, concerning the MrRegNet-SS results on the 2D dataset, the utilisation of a scaling and squaring layer with $t = 5$ leads to slower training and inference speed. This highlights one of the advantages of our model: good diffeomorphic performance is achieved without the requirement of the scaling and squaring layer, thereby maintaining higher efficiency.

5.4 Discussion and Conclusions

To achieve an end-to-end image registration that is able to cope with large deformations, a DCNN based multi-resolution registration framework is proposed in this chapter. It learns a residual displacement field in each resolution. By using a smoothness term of equal weights on all residual displacement fields, pixels at each resolution only move small distances in their own scales. This design preserves the properties of diffeomorphic deformation. The evaluation results show that the proposed method can achieve better performance than the commonly used non-learning-based method Demons and a well-known learning-based method VoxelMorph on the 2D brain MRI dataset with large image deformations. The method is also able to achieve high quality registration results on the 3D brain MRI dataset with multi-class masks.

Based on comprehensive experimental evaluations, it is concluded that the proposed MrRegNet-G (mask) is the most preferable method in terms of global alignment, local alignment and diffeomorphic properties. This is attributed to the multi-resolution residual displacement field learning and global NCC similarity measure combined with a mask-guided loss.

Moreover, the proposed image registration model can be utilised as an image segmentation tool. The segmentation mask of a template image (source image) can be warped to the target image space using the estimated displacement field. The inclusion of the mask guided loss significantly improves the registration performance in the masked regions, which can further improve the segmentation accuracy. Therefore, the next chapter will introduce a novel semi-supervised segmentation framework by combining both a deep-learning based segmentation model and the proposed MrRegNet-G (mask). This framework is trained on a small amount of annotated images and a large amount of unannotated images to iteratively improve the segmentation and registration models.

Chapter 6

Integrated Image Segmentation and Registration for Semi-supervised Learning

6.1 Introduction

In chapter 5, an end-to-end unsupervised image registration framework is presented, which can be applied to both 2D and 3D images. The framework estimates a displacement field that warps the source image to the target image. Consequently, this displacement field can also be used for generating a segmentation mask of the target image by warping the mask of the source image. By incorporating a mask-guided loss term, the model learning process is directed to prioritise the masked region, resulting in improved registration and segmentation performance for that specific area.

In chapter 4, a method based on ensemble techniques is proposed to generate pseudo-masks, and a semi-supervised image segmentation method is developed.

Based on the ideas from both chapter 4 and 5, in this chapter, the image registration model (chapter 5) is combined with a CNN-based image segmentation model to produce pseudo-masks of unannotated images, aiming to

simultaneously improve the performance of both models in semi-supervised learning.

Before the widespread adoption of deep learning, it had been discovered by researchers that image registration and image segmentation could mutually enhance the results of both models by providing valuable information to each other. In 2001, Yezzi et al. introduced the first joint framework for image registration and segmentation [152], which employed active contours to simultaneously segment and register features across multiple images. Subsequently, several non-learning-based methods were proposed, such as grow-cut based [153], Bayesian based [154] and Markov random field based [155].

However, all of those methods operate on individual pairs of images, resulting in high computational complexity. In 2019, Xu and Niethammer introduced DeepAtlas, a deep learning-based framework that jointly integrates image registration and segmentation [21]. They combined the registration network and segmentation network, connecting them with an anatomy similarity loss that assesses the similarity of the generated masks produced by both models. Nonetheless, the performance heavily relies on the quality of the labels generated by each model. A single incorrect result can have adverse effects on the learning of both models, resulting in a gradual decline in performance over time.

In this chapter, to address the problem mentioned above, a novel component called "soft pseudo-mask generation" is added to the joint image registration and segmentation framework. It includes an automatic evaluation mechanism to measure the quality of segmentation results during training. Similar to DeepAtlas, this framework also enables semi-supervised learning, where the model is trained using a few annotated data and a large amount of unannotated data. However, the difference is that the proposed framework does not train the model using all the unannotated data directly in a single run. Instead, it refines the segmentation and registration model iteratively. In each iteration of the training phase, the framework fuses the masks of unanno-

tated images generated by both models. Each of these fused masks provides a pixel-wise confidence map to guide improving the performance of both models. By iteratively training the model using these confidence masks, it prevents the model from learning incorrect information. This approach ensures a steady improvement in training and enhances the overall performance of both models. The proposed framework has been evaluated on a 2D brain MRI dataset. The results show that this method can significantly improve the performance of image registration and image segmentation based on very few annotated images.

The remaining sections of this chapter are organised as follows. Section 6.2 presents the overall architecture of the joint training framework and provides detailed information on the soft pseudo-mask generation element. Section 6.3 outlines the experimental details, including dataset and experimental settings. It also shows the results of all the experiments. Finally, Section 6.4 summarises the conclusions drawn from this chapter and provides a discussion of the findings.

6.2 Methodology

6.2.1 Framework Architecture

Figure 6.1 illustrates the proposed joint training framework for image segmentation and registration. The two models are trained in an iterative manner. Firstly, the image segmentation model is trained on all annotated images to create an initial model called "Segmentation model *0*". Same as described in chapter 5, the image registration model is initially trained on the entire training dataset, which includes both annotated and unannotated images, using unsupervised learning without the use of any masks, resulting in "Registration model *0*". Then, both models generate several candidate pseudo-masks for each of the unannotated images. The soft pseudo-mask generation component

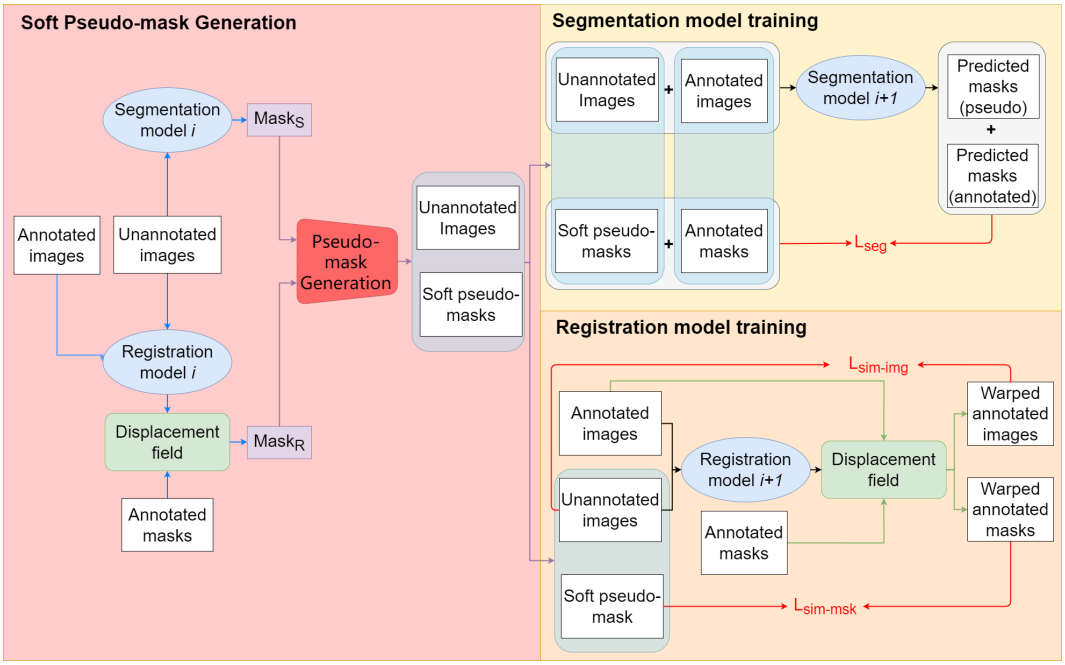

Figure 6.1: *Overview of the proposed joint training framework for one training iteration. The framework consists of three components: Soft pseudo-mask generation, Segmentation model training, and Registration model training. The Soft pseudo-mask generation component combined the masks generated by the segmentation and registration models for unannotated images into soft pseudo-masks. Subsequently, a new training set is formed by combining annotated images and the unannotated images with pseudo-masks to refine both the segmentation and registration models.*

combines these pseudo-masks for the next iteration of training. Intuitively, the segmentation model and registration model iteratively improve each other via the gradually improved psuedo-masks of the unannotated images. Detailed information of each component is provided below.

Image Segmentation Model

The image segmentation model used in this chapter is a modified version of the widely used medical image segmentation network, U-net[3]. It uses a multi-resolution encoder-decoder structure, with 5 levels of resolutions in both the encoder and decoder. This allows the model to effectively extract features and generate segmentation predictions for the input image. In our proposed method, residual blocks are added to the conventional U-net for a more efficient feature learning and faster model training.

In the encoder, each network level consists of a combination of a residual

block [83] and a max-pooling layer. The residual block includes two convolutional layers (Conv) with a kernel size of 3×3. After each convolutional layer, there is a batch normalisation layer (BN) [156] and a rectified linear unit (ReLU) [137]. To achieve deep feature extraction, the first Conv doubles the number of channels in the feature map (16 for the very beginning Conv which input is the image itself). The second Conv maintains the same channel size. As a result, the output feature map have twice the channel size compared to the input feature map. Additionally, to achieve residual learning, the input feature map of the residual block is added as a residual to the output feature map of the second BN. Then, the second ReLU is applied. The max-pooling layer, with a stride of 2, down-samples the feature map output from the residual block, reducing its size by half. The final high resolution level in the encoder does not include a max-pooling layer. Instead, the feature map obtained from the residual block is directly passed to the decoder without any down-sampling.

In the decoder, each network level consists of a deconvolutional layer (Deconv) with a 3×3 kernel size, followed by a ReLU activation operator, and a residual block. The deconvolutional layer serves to decode and up-sample the feature map from the previous layer, resulting in a feature map that is double the size and has half the number of channels. Next, a skip connection is applied, where the feature map from the corresponding level in the encoder is concatenated with the up-sampled feature map in the channel dimension. The combined feature map is then activated by the ReLU function and passed to the residual block. The structure of the residual block in the decoder is similar to that in the encoder, but with a halved channel size instead of doubling it. This ensures that the output feature map matches the desired size and channel size of the Deconv. Finally, at the end of the decoder (the highest resolution level), a Conv with a kernel size of 1 is used to adjust the size of the feature map output from the residual block to match the desired size of the segmentation mask.

Image Registration Model

The image registration model utilised in this chapter is the model proposed in chapter 5, named MrRegNet-G. This model adopts a multi-resolution structure of residual displacement field, enabling effective image registration for images with both small and large deformations. It is designed to conduct image registration using unsupervised learning. Additionally, the MrRegNet-G (mask) method leverages the corresponding masks and employs a mask-guided loss to enhance attention and improve the registration performance specifically within the masked regions. For more detailed information about the architecture of MrRegNet-G, please refer to section 5.2.1. It is worth noting that, similar to chapter 5, the registration models were trained using random paired source and target images to improve the capability of model generalisation.

Soft Pseudo-masks Generation Block

Besides the above two models, another key component is the soft pseudo-mask generation block. In DeepAtlas [21], the image segmentation model and the image registration model are jointly trained, leveraging the pseudo-masks generated by each other for the unannotated images. However, it is important to note that this mutual improvement mechanism can potentially lead to sub-optimal learning outcomes when the generated results are incorrect.

To address this issue, an iterative joint training process is adopted in this chapter. A soft pseudo-mask generation block is introduced to combine the pseudo-masks generated by both models at each iteration and produce a final soft pseudo-mask for each unannotated image. Here "soft" means that instead of integer labelled mask, the pixel values in the soft mask range between 0 and 1, representing confident scores. This enables the creation of a training set with pseudo masks specifically tailored for the unannotated images. Subsequently, this set is used to train both models in the next iteration.

Through this iterative process, the quality of the soft pseudo-masks are

also improved, mitigating the risk of the image segmentation model and the image registration model learning incorrect information from inaccurate masks.

6.2.2 Soft Pseudo-mask Generation Strategy

Both the image segmentation and the image registration model have the ability to generate pseudo-masks by using the initially trained models or the updated models from the previous iteration. For the image segmentation model, the pseudo-mask is generated by inputting an unannotated image, and the model outputs a predicted mask that can be considered as a pseudo-mask. As for the image registration model, once trained, it can generate a displacement field by inputting paired source (annotated) and target (unannotated) images. This displacement field represents the deformation mapping from the source to the target image. By warping the mask of the source image using this displacement field, a pseudo-mask for the target image can be generated. The soft pseudo-mask generation block utilises the pseudo-masks generated by both models to generate "soft pseudo-masks" that are used to refine the models.

To increase the reliability and quality of the pseudo mask, instead of generating one mask from each model, a set of pseudo-masks are generated from each of the segmentation and registration models. The process to achieve this is illustrated in Fig. 6.2.

For the segmentation model, it applies test time augmentation to each of the unannotated images. During test time augmentation, each unannotated image is augmented using a randomly selected augmentation method (i.e. rotation or shift in this case) with random parameters such as rotation angle and shift distance. The segmentation model then generates a probability map for each class for the augmented image and transform them back to the original image space. This process is repeated for a total of $N-1$ times, resulting in N probability maps comprising one map for the original image and $N-1$ maps for the augmented images.

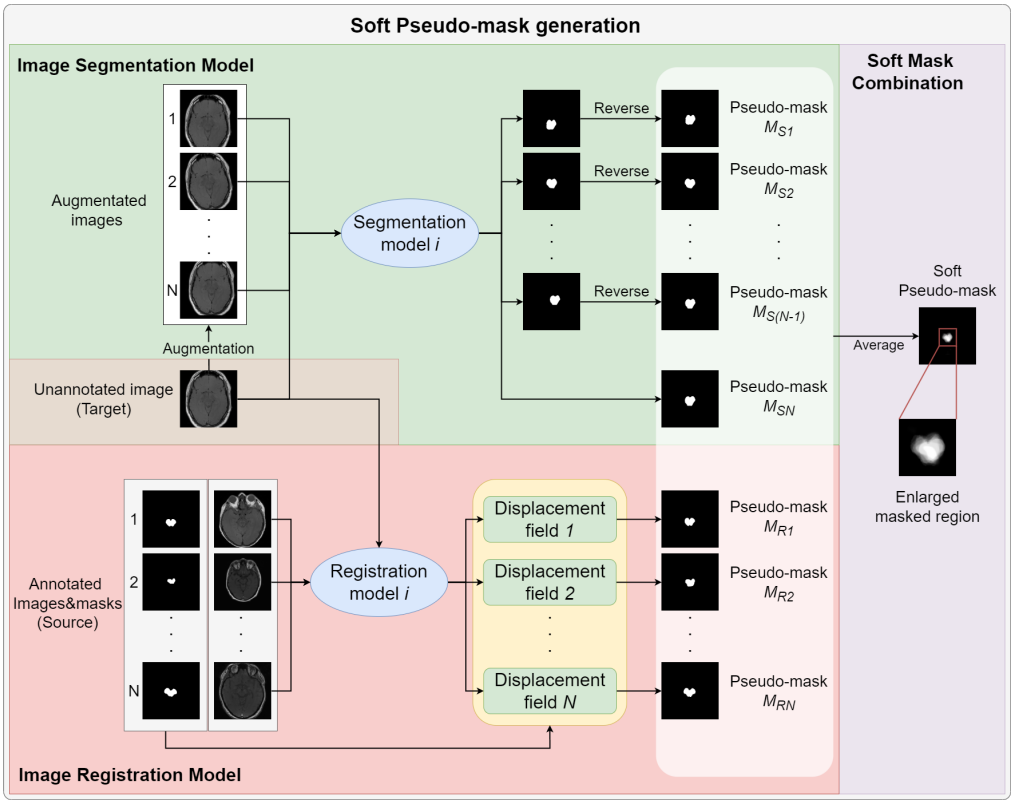

Figure 6.2: *Soft pseudo-mask generation by image segmentation and image registration models for each unannotated image. The image segmentation model generates N probabilistic pseudo-masks by applying test-time data augmentation. The image registration model generates a displacement field that represents the mapping from each template annotated image to the unannotated image. This displacement field is then used to warp the annotated mask, resulting in a pseudo-mask for the unannotated image. The averaged map of all the $2N$ pseudo masks is used as the final soft pseudo mask.*

Similarly, for each unannotated image, the image registration model also generates N masks to ensure a balanced contribution compared to the segmentation model. The image registration model generates the pseudo masks for the unannotated image by warping the masks of N annotated source images. This means that by utilising multiple different source images and their corresponding masks, the model can generate multiple masks for the same unannotated image.

Finally, $2N$ pseudo masks are generated for each unannotated image by both the segmentation model and registration model. The averaged map of all the $2N$ pseudo masks is used as the final soft pseudo mask. The value range of the pseudo mask is between 0 and 1. A higher value indicates a higher

confidence of a pixel belonging to the assigned class. It can be generalised to the case of multiple classes: one soft pseudo mask for each class. Specifically, the image segmentation and the image registration models generate $2N$ multi-channel probability maps for an unannotated image. Each channel in the probability maps corresponds to a specific class, and the pixel value represents the likelihood that the location belongs to that particular class.

When dealing with medical image segmentation, it is natural to encounter uncertainty as various plausible segmentation hypotheses can emerge for a given image. Recent studies [157] [158] have provided evidence that as the number of independent annotations reaches a specific threshold, the variability in segmentation tends to be stabilised. This discovery suggests that if a sufficiently large group of physicians is involved, they could potentially encompass the entire range of possible segmentation. Therefore, the "soft mask" approach involves the utilisation of multiple pseudo-masks to simulate the annotation of a single image by multiple annotators.

6.2.3 Model Training

To train the proposed framework, an iterative training approach is employed to gradually refine both the segmentation and registration models, as show in figure 6.1. To start at iteration 0, the segmentation model is trained on all annotated images as the initial model. On the other hand, the registration model undergoes unsupervised training on all images without utilising any masks. These initial models are referred to as pre-trained models. Subsequently, the soft pseudo-mask generation block utilises these pre-trained models to generate a soft pseudo-mask for each of the unannotated images. From that point onward, the unannotated images with their corresponding pseudo-masks are combined with the annotated images to form a new training set, which participates in the training process of the following iteration.

Moving on to the next iteration (iteration 1), the pre-trained models ob-

tained from the previous iteration (iteration 0) are refined using a smaller learning rate and a reduced number of training epochs using the new training set generated from iteration 0. Once both models are updated, the aforementioned steps are repeated to generate a new set of pseudo masks for the unannotated images to form an updated training set, which is used in the subsequent iteration.

The loss function for model training is a crucial component of any deep learning model. In this proposed framework, the utilisation of the soft pseudo masks for training is the key to successfully improve both models. Conventional loss functions for segmentation are Dice coefficient and cross entropy, which are calculated based on integer labelled masks. In contrast, the soft Dice loss function is utilised here to optimise the models. Additionally, following the approach outlined in the V-Net [71], squaring is applied to the denominator of the loss function to create a smoother landscape for faster convergence. The equation for the soft Dice loss function is expressed as follows:

$$L_{dsc}(y', y) = 1 - \frac{2\sum_p^{\Omega}({y'}_p y_p)}{\sum_p^{\Omega}({y'}_p^2) + \sum_p^{\Omega}(y_p^2)} \tag{6.1}$$

where p indicates the index of pixels in the whole image Ω. y and y' represent the annotated mask (or soft pseudo mask) and the predicted probability map, respectively.

Furthermore, the image registration model integrates a similarity loss and a smoothness term, following the approach described in chapter 5. Specifically, the global nearest cross-correlation is employed as the similarity loss, and the smoothness term is defined by an $L2$ regularisation term for the displacement field. The definition of these terms can be found in Eq. (5.3) and Eq. (5.5), respectively.

In summary, the image segmentation model and registration model are trained by optimising the subjective functions in Eq. (6.2) and Eq. (6.3)

respectively.

$$arg\ min\, L_{dsc}(y', y) \tag{6.2}$$

$$\underset{D}{arg\ min}\, \frac{1}{K} \sum_{i=1}^{K} (L_{sim}(f_{D_i}(x_S), x_T) + L_{dsc}(f_{D_i}(y_S), y_T) + \lambda L_{smooth}(D_i)) \tag{6.3}$$

where D represents the displacement field, and D_i denotes the final displacement field at each resolution out of K resolutions. L_{gncc} and L_{smooth} are the similarity loss and smoothness regularisation loss, respectively. Additionally, x_S and x_T indicate the source and target images, while y_S and y_T represent the source and target masks, respectively. Moreover, $f_D(x_S)$ and $f_D(y_S)$ signify the warped source image and the warped source mask respectively. The model is updated by maximising the similarity between the warped source image and the target image, maximising the soft Dice score between the warped source mask and the target mask, and minimising the $L2$ norm of the displacement field.

6.2.4 Model Inference

After finishing the joint model training, the image segmentation and image registration models from the final iteration can be used to segment unseen images. Similar to other methods, the segmentation model has the ability to directly segment an unseen image. Meanwhile, the image registration model can utilise any specified or randomly selected annotated images as the source image, while the unseen image serves as the target image. The estimated displacement field can be used to map the source image to the target image, which is then used to warp the source mask to the target space. Consequently, the warped mask of the source image can be considered as the segmentation mask for the target image.

A single mask generated from either the segmentation model or the regis-

tration model may not be sufficiently accurate. One key advantage of the proposed joint framework is that the final mask could be generated by combining several outputs from both models using the "Soft Pseudo-mask Generation" component. Following the same process as illustrated in Fig. 6.2, by applying test time data augmentation to the segmentation model and selecting several source images for the registration model, a soft pseudo mask can be generated for a given unseen image. Then, the "argmax" operation is applied to the soft mask to generate the final combined mask. This combined mask represents an aggregated decision that takes the advantages of both models, leading to improved performance compared to using either the segmentation model or the registration model (see section 6.3.3 for more details).

Moreover, the soft pseudo-mask can serve as a confidence map (as illustrated in Fig. 6.3) based on the aggregated decision from multiple outputs. This pixel-wise confidence map can be utilised for various purposes, such as uncertainty estimation, segmentation quality assessment, and other downstream tasks. By analysing the confidence values assigned to each pixel or region in the pseudo-mask, one can gain insights into the uncertainty associated with the predictions. This information is valuable for understanding the reliability of the model's outputs and can guide decision-making and further analysis. Figure 6.3 shows an example, including the original image, the corresponding ground truth mask, the generated soft mask by the trained framework and the final combined mask by applying argmax to the soft mask.

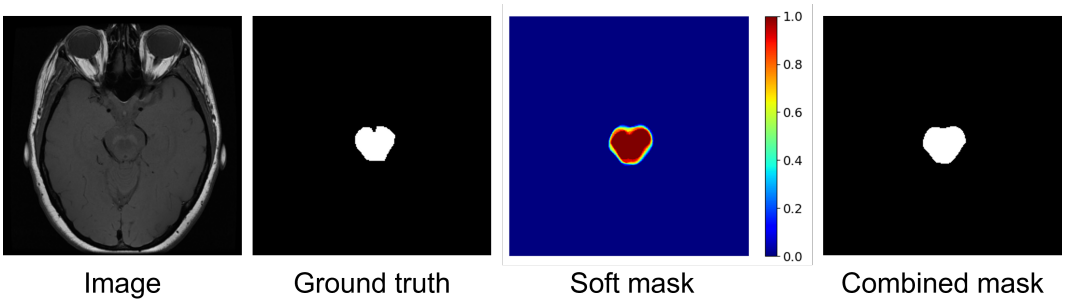

Figure 6.3: *Left to right: an example of an input image, the corresponding ground truth mask, the heatmap of the soft pseudo-mask and the final combined mask by applying argmax to the soft mask.*

6.3 Method Evaluation

6.3.1 Dataset

The evaluation of the proposed method was conducted on the same 2D dataset used in chapter 5. This dataset comprises 2D brain MRI slices from hundreds of subjects. These images exhibit significant intensity and geometric variations, as they were from different institutions and different scanners. As a result, segmenting this dataset poses a considerable challenge.

The dataset consists a total of 820 images, which were divided into training and test sets using an approximate 80-20% split (i.e., 620 for training and 200 for testing). Furthermore, a small portion (20 images) of the training set was selected for validation purposes. Therefore, the final distribution of images resulted in 600 images for training, 20 images for validation, and 200 images for testing.

Furthermore, in response to the varied image sizes and intensities in this dataset, pre-processing was applied to the images. Specifically, the images were resized to 256×256 pixels, and their intensities were normalised to the range of 0 to 1 using min-max normalisation. These pre-processed images were used consistently in all experiments described in this chapter.

6.3.2 Experimental Design

The main objective of this study is to experimentally evaluate the effectiveness of the proposed joint training framework for semi-supervised image segmentation. To achieve this, three experiments were conducted. In the few-shot learning scenario, only five (around 1%) annotated images were utilised, while for conventional semi-supervised learning, 10% (60) and 30% (180) of the annotated images were used, respectively. These experiments aimed to assess the performance and efficacy of the proposed method under different levels of annotated data availability.

To ensure fair comparisons, all experiments utilised the same unsupervised image registration pre-trained model at iteration 0. Additionally, each experiment using the proposed framework underwent a total of 10 iterations. This allowed for a closer examination of how the models are evolved and improved over time. Furthermore, to establish an upper bound, the image registration and image segmentation models were trained using all available annotated data in a fully supervised manner. This upper bound served as a reference point, providing insights into the maximum potential performance achievable with the given dataset and models.

Parameter Settings

All three joint training models share the same parameter settings for the image segmentation and image registration networks. Specifically, for the image segmentation network, the learning rate of the pre-trained model in iteration 0 was set to 0.0001, and the number of training epochs was set to 500. In the subsequent iterations, the learning rate was reduced to 0.00001, and the training epoch was decreased to 100. The batch size for both models were set to 5 according to the GPU memory limitations.

For the image registration network, the parameter settings align with those in chapter 5. The learning rate was set to 0.001, and the training epoch was set to 200 for the pre-trained model in iteration 0. The weights of the smoothness term, from the lowest to the highest resolution, were set as follows: 128, 64, 32, 16, and 8. In the remaining iterations, the learning rate was reduced to 0.0001, and the training epoch was set to 50 for fine-tuning the network. The weights of the smoothness term remain unchanged.

To expedite the computational process of the soft pseudo-mask generation block, the number of pseudo-masks generated by each network (N) was set to 5 for all experiments. In cases the model trained on more than 5 annotated images, the image registration model randomly selects 5 annotated images as the source images to generate pseudo-masks for each unannotated image.

Similarly, the segmentation model generates 5 pseudo-masks, comprising one segmentation result on the original input image and four segmentation results on augmented images. As a result, a total of 10 pseudo-masks were produced from both models for soft pseudo-mask generation.

Furthermore, to ensure a fair comparison, the fully-supervised image segmentation and image registration models share the same parameter settings as the pre-trained models. These parameter settings and pseudo-mask generation approach were devised to optimise computational efficiency while still providing reliable results for the soft pseudo-mask generation process.

6.3.3 Results

Semi-supervised Image Segmentation and Registration Results

The evaluation results of the segmentation and registration model are reported separately. For image segmentation, the Dice coefficient (DSC) is utilised as the evaluation metric, which is commonly employed in image segmentation tasks. As for the image registration model, similar to chapter 5, the overall and local image registration performance is evaluated using the normalised cross-correlation (NCC) and Dice coefficient, respectively. However, unlike chapter 5, this chapter focuses specifically on image segmentation performance, therefore does not employ the Jacobian determinant as an evaluation metric. Moreover, the DSC of the combined mask is reported as one of the outputs of the proposed joint training framework.

Table 6.1 presents the results of the joint training methods. The fully-supervised image segmentation and image registration (i.e. MrRegNet-G (mask)) models are denoted as "F-100%", and their results are listed in the same row. Joint-1%, Joint-10%, Joint-20%, Joint-30%, Joint-40%, Joint-50% and Joint-60% are the joint models trained using the annotated/unannotated ratios of 1%/99%, 10%/90%, 20%/80%, 30%/70%, 40%/60%, 50%/50% and 60%/40%, respectively. "B" in table 6.1 refers to the baseline result of each model, which

Table 6.1: Numerical results for the fully-supervised models and the proposed methods are presented. The fully-supervised image segmentation and image registration models are denoted as "F-100%". The "B" refers to the baseline result of each model, which corresponds to the performance of the pre-trained model at iteration 0. "–" indicates that the results are the same as Joint-1%B as they use the same pre-trained model. The DSC values are reported for segmentation models, registration models and the combined mask, while NCC is used only for the registration model. The reported values are presented as the mean ± standard deviation.

Method	**Segmentation**	**Registration**		**Combined Mask**
	DSC	**NCC**	**DSC**	**DSC**
F-100%	0.93±0.02	0.76±0.04	0.86±0.06	n/a
Joint-1%B	0.58±0.27	0.77±0.04	0.77±0.10	n/a
Joint-1%	**0.84±0.05**	0.81±0.03	0.83±0.06	**0.84±0.05**
Joint-10%B	0.83±0.15	–	–	n/a
Joint-10%	0.86±0.06	0.81±0.03	0.86±0.06	**0.87±0.04**
Joint-20%B	0.88±0.06	–	–	n/a
Joint-20%	0.88±0.05	0.80±0.03	0.87±0.05	**0.89±0.04**
Joint-30%B	0.91±0.03	–	–	n/a
Joint-30%	0.89±0.04	0.80±0.03	0.87±0.05	**0.90±0.03**
Joint-40%B	0.91±0.06	–	–	n/a
Joint-40%	0.91±0.04	0.80±0.03	0.87±0.05	**0.91±0.03**
Joint-50%B	0.92±0.05	–	–	n/a
Joint-50%	0.92±0.05	0.80±0.03	0.87±0.05	**0.92±0.03**
Joint-60%B	0.92±0.03	–	–	n/a
Joint-60%	**0.92±0.03**	0.80±0.03	0.86±0.06	**0.92±0.03**

is the performance of the pre-trained model at iteration 0.

First of all, it can be seen from table 6.1 that the joint model training can improve the segmentation performance significantly compared to the pre-trained model when a small amount of annotated images is used. Specifically, with just 5 (1%) annotated images, a remarkable improvement of the segmentation model is observed, with an increase from 0.58 to 0.84. Similarly, for Joint-10%, which involves 60 annotated images, there is an improvement of more than 3%, resulting in a performance increase from 0.83 to 0.86. However, as the amount of annotated data increases, the performance improvement of the segmentation model becomes neglectable or even with a slight decrease (i.e. Joint-30%). It is hypothesised that this phenomenon is due to a reduced contribution from the unannotated images when the number of annotated images is sufficiently large. Also, when the performance of the image registration

model reaches certain level, the quality of the soft pseudo labels can not be improved further.

Secondly, for the image registration model, it shows a consistent improvement compared to the baseline model in terms of the DSC values. As the number of annotated data increases, the registration model shows a steady improvement, and comparable to the performance of the fully-supervised learning when more than 10% of annotated data are utilised. Furthermore, the iteratively trained image registration model achieved a consistent and better performance than the fully-supervised model using the NCC measure. The NCC values of all the joint training models are around 0.8, indicating a good global image alignment. This is even better than the fully supervised learning method F-100% (0.76). This improvement is contributed by the soft masks for guiding the image registration's attention. Instead of using binary masks by the F-100% method, the joint model ultilises soft masks. It prevents the model from excessively focusing on the masked region, which leads to an improved global alignment while still maintaining a good local segmentation quality.

Finally, the combined mask generated from both models achieved the highest segmentation performance with smaller standard deviation in all tests, compared to the individual segmentation and registration model. This finding suggests that the soft pseudo-mask generation block not only helps to improve the joint model training iteratively, but is also capable of producing a high quality segmentation result in the model inference process.

Performance Evolution in the Training Process

The train process of the joint model was also monitored by reporting the segmentation accuracy (DSC value) at each iteration. The DSC values of the segmentation model, registration model and the combined mask for each iteration using different percentages of annotated data are plotted in Figure 6.4.

It can be seen that both the image segmentation and registration models

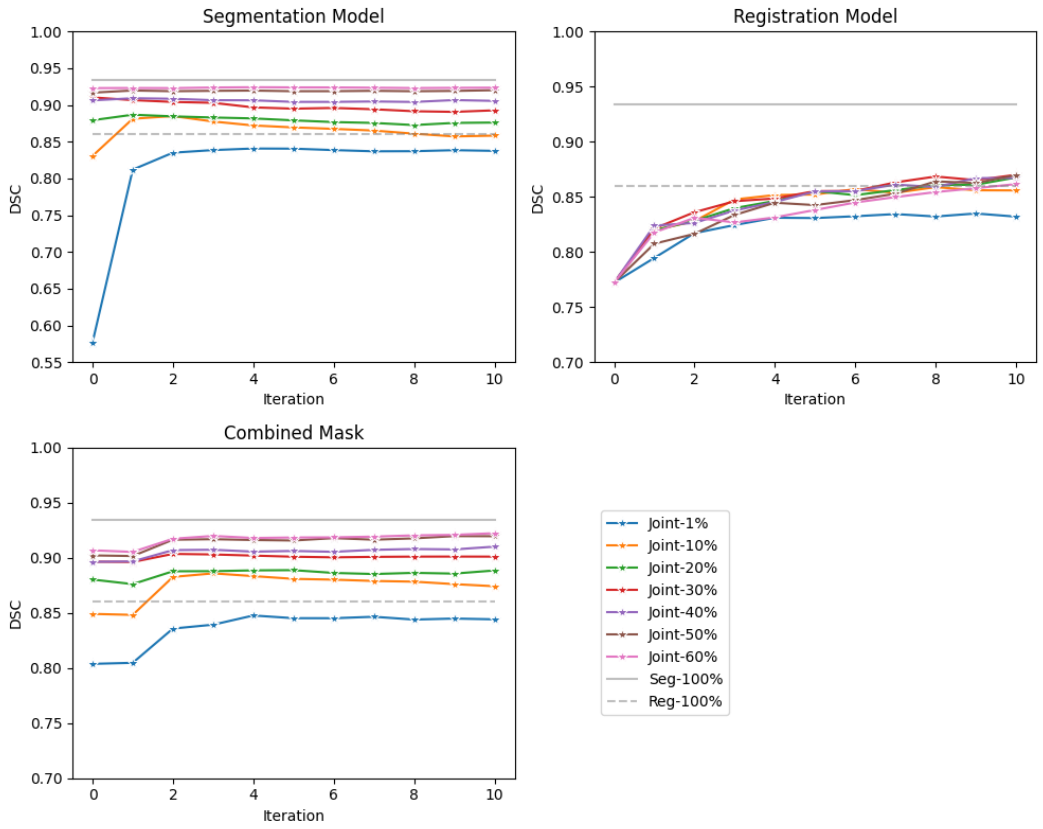

Figure 6.4: *The performance of various methods across different iterations on the test set. The horizontal axis corresponds to the iteration number, where iteration 0 represents the pre-trained model. The vertical axis represents the evaluation score for the respective models. The proposed joint training framework is represented as "Joint". The notation "-n%" indicates the percentage of the number of annotated images used for model training. "Seg-100%" and "Reg-100%" indicate the fully-supervised image segmentation model and the image registration model, respectively. They are represented by two lines in each plots as baseline.*

undergo significant improvements through iterative training for Joint-1% and Joint-10% models. By increasing the number of annotated data, the performance improvement of the segmentation model is less significant. When using 10%(60) and 20%(120) annotated data, the performance improvement is only observable in the first few joint training iterations. However, when the amount of annotated data is increased to 30%(180) above, the segmentation model is no longer improved through joint training. This is due to the number of annotated images are sufficiently large to guide the model training. The registration models of all cases are consistently improved throughout the training process, and become stabilised after 8 iterations.

The second row in Fig. 6.5 displays the segmentation performance of the

combined mask produced by the "soft pseudo mask generation" component . It is evident that the performance of the combined mask is better than the two individual models in all test cases. It can also be observed that after reaching the peak, the changes in DSC remain relatively small in the subsequent iterations, and there is no significant decrease compared to the image segmentation model. This indicates that the proposed method can effectively combine the advantages of both models to achieve a robust decision making.

Visualisation of the Soft Pseudo Mask

More qualitative analysis was conducted to observe the improvement of soft pseudo masks during training. Fig. 6.5 shows a set of generated pseudo masks generated by both models in different iterations using the Joint-1% model. The corresponding segmentation outputs from each individual model (segmentation and registration) are also shown in Fig. 6.5 for comparison.

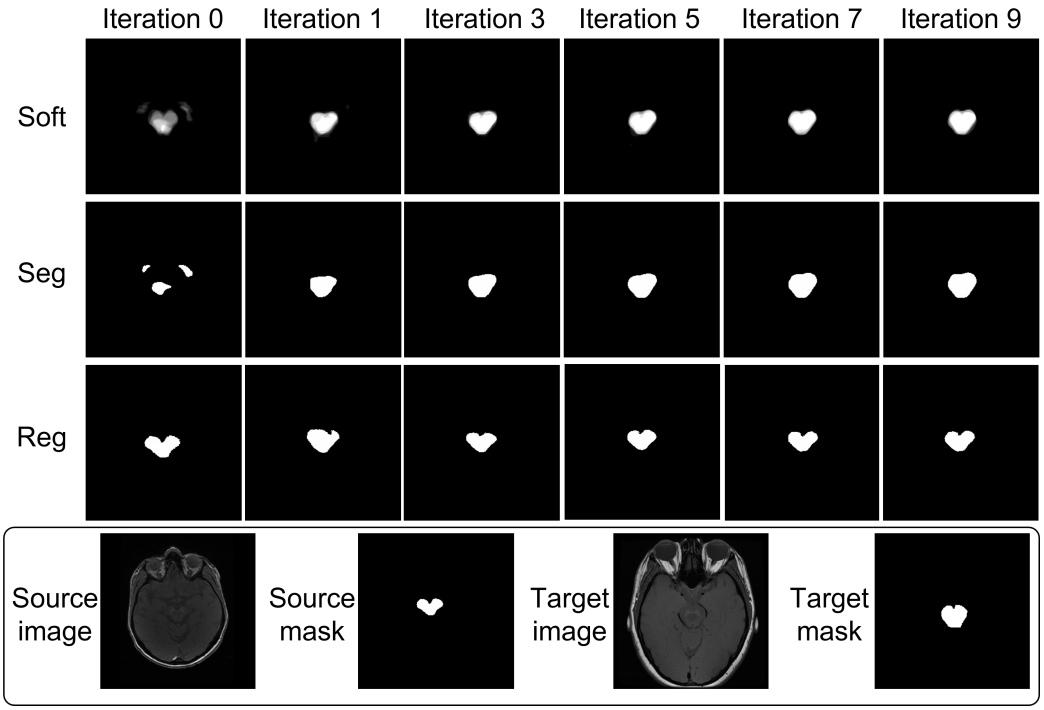

Figure 6.5: *Visualisation results of an image that participated in training as an unannotated image in Joint-1%. The soft pseudo-masks (Soft), segmentation model results (Seg) and the registration model results (Reg) at different iterations are provided. At the bottom, the source image, the source image mask, the target image and the annotated target image are presented.*

The images in the first column (iteration 0) demonstrate that the pre-trained segmentation model produces significant errors in the segmentation result, while the pre-trained registration model (without mask guidance) can only achieve a coarse alignment in the local region. Furthermore, the soft pseudo-mask contains a very small region of agreement.

As the training progresses from iteration 1 onward, more pixels in the soft pseudo-mask become more confident (larger brighter regions indicate larger overlapped mask regions). The improved soft mask helps both the segmentation and registration models to achieve better performance iteratively. The joint model is converged quickly at around 7 iterations. The final soft mask can be used as a confidence map indicating pixel-wise segmentation uncertainty.

Computational Time

In addition to the segmentation quality, it is important to further discuss the model training and inference time in detail. All the computational time reported below was using a GPU server an Intel E5-2620 v4 CPU running at 2.1GHz, and a NVIDIA 1080Ti GPU with 11GB memory. The code was implemented using the PyTorch deep learning framework in Python.

For the pre-trained segmentation model, the average training time was about 84 images per second. Due to the variations of the training size, the training time differs for different data percentages. For 1%(5), 10%(60), 20%(120), and 30%(180) images, the training time for 500 epochs were 29, 357, 714, 1071 seconds, respectively. In each of the subsequent iterations, a total of 600 images were trained with 100 epochs, taking approximately 12 minutes per iteration.

Regarding the registration model, both the pre-training and iterative training models were trained for about 38 pairs of images per second. Therefore, the time required for the pre-training model to train 200 epochs on all 600 images was approximately 50 minutes. For each of the subsequently iterations, the model was trained with 50 epochs, which resulted in about 13 minutes for each iteration. In each iteration, additional time was required for

soft pseudo-mask generation, which took about 6 minutes for 600 images. In total, a complete training session for the joint model required about 15 hours.

As for model inference, the prediction time for the segmentation model per image is about 0.007 seconds, while the prediction time for the registration model per image is about 0.04 seconds. The final prediction based on the soft pseudo-mask requires data augmentation and multiple inferences by each model on the same image, resulting in approximately 0.7 seconds per image.

In summary, although the joint model training takes 15 hours, once the model is trained, it only takes a fraction of second to produce a segmentation result.

6.4 Discussion and Conclusions

This chapter has introduced a framework that enables semi-supervised joint training of an image segmentation model and an image registration model. The framework adopts an iterative approach, starting with the training of a segmentation model using a small amount of annotated data, and an unsupervised image registration model trained on both annotated and unannotated data. Subsequently, the iterative training process involves passing through a pseudo-mask generation block, which generates soft pseudo-masks used to refine both models.

The experimental results demonstrate that this approach enables the two models to mutually improve each other's performance, achieving accurate image alignment and segmentation even with limited annotated data (1%). Notably, the utilisation of the registration method in this framework enhances the ability to preserve the anatomical structural information. Consequently, the jointly trained segmentation model also inherits this crucial feature, which holds significant importance, especially in the domain of medical imaging.

Additionally, the results indicate that the soft pseudo-mask generation block successfully generates soft pseudo-masks, which are then utilised to refine

both models. Furthermore, the pseudo-masks generated from both models are fused together to produce a final segmentation result in the model inference process. Moreover, the soft mask can be interpreted as a confidence map which is useful in various downstream tasks, such as quality control, uncertainty estimation, etc. The performance of the proposed framework could be further improved by increasing the number of augmented images (segmentation model) and the number of template source images (registration model) for soft mask generation.

With this, the technical components of the thesis are concluded, and the next chapter will provide a summary and discussion of the entire thesis.

Chapter 7

Conclusions and Future Work

This chapter is the last part of the thesis and aims to provide a summary of the discoveries, contributions, limitations, and potential areas for further investigations.

7.1 Conclusions and Contributions

As mentioned in **chapter 1**, the goal of this research work is to create an efficient image segmentation solution that is completely automated to segment new images, and with reduced efforts from clinical experts to develop such a solution. To accomplish this, semi-supervised machine learning strategy was considered as the main area of interest. A fully automated segmentation model can be trained by using a small number of annotated data and a large amount of unannotated data. In order to achieve this goal, the following objectives were identified:

- Developing an efficient annotation tool that enables experts to annotate a small number of data samples.

- Developing a method that can be trained using a limited set of annotated images and a substantial amount of unannotated images, aiming to achieve comparable performance to a fully supervised method.

In **chapter 2**, a literature review was provided in the context of medical image segmentation. The advancement from traditional image segmentation methods to deep learning based methods was discussed, with a particular attention on the widely used U-Net method for medical images. Additionally, a review of open-source manual software was conducted, revealing that most existing interactive segmentation software is primarily effective for 2D natural images and performs poorly when applied to medical images. Furthermore, many of these software lack the ability to provide multi-label image annotation. The annotation approach employed by several software tools involves contour labelling and region filling, which is a time-consuming process.

More importantly, an overview of semi-supervised learning methods is presented, including commonly employed methods such as data augmentation, consistency regularisation, and pseudo-labelling. It is observed that pseudo-labelling-based methods can be directly and efficiently applied to image segmentation tasks, although the challenge remains on preventing the model from learning erroneous information from imperfect pseudo-labels. Furthermore, the widely used image registration methods in medical images were introduced, with a focus on its application to the field of medical image segmentation. A small number of methods that combine image segmentation and image registration were discussed. It is discovered that image segmentation and image registration models can benefit each other if they are combined together.

In **chapter 3**, a CRF-based interactive segmentation software was developed. This software is capable of performing multi-label segmentation on both 2D and 3D medical images. Users can easily use this software without a complicated training process. Additionally, the software allows users to refine inaccurate segmentation regions to improve accuracy. In the case of 3D data, segmentation can be performed slice by slice, and users can modify annotations from different views. The software can also recommend the best slice to annotate based on information entropy, streamlining the segmentation process and reducing the required time and effort.

The key contributions of this study can be summarised as below: (1) an open-source interactive image segmentation software for both 2D and 3D multi-label medical image segmentation; (2) a novel slice recommendation function for 3D images to improve segmentation efficiency; (3) "one size for all" parameter setting for different image modalities and dimensions. This work is published as a journal paper, and the details can be found in Publication 1.

After addressing the manual annotation challenge, in **chapter 4**, a semi-supervised learning framework is proposed, leveraging both annotated and unannotated data to train the segmentation model. Unlike labels for classification and regression problems, image segmentation labels (also called masks) have a higher level of complexity. Hence, a self-learning and pseudo-labelling approach is adopted to achieve semi-supervised learning in this proposed method. The key to effectively using pseudo-labels is controlling their quality. To achieve this, an ensemble technique is introduced, involving iteratively training a small set of models to gradually improve the quality of pseudo-masks and enhance the model performance. Evaluation on a public 2D skin lesion dataset demonstrates that this method achieves state-of-the-art performance in semi-supervised image segmentation.

The main contributions of the semi-supervised learning framework are: (1) a new generic end-to-end semi-supervised learning framework; (2) an effective ensemble technique to control the quality of pseudo-labels for semi-supervised image segmentation model training;(3) state-of-the-art performance on a public dataset. The paper is published in IEEE-ISBI conference with detailed information listed in Publication 2.

Chapter 5 explored the use of image registration method to achieve image segmentation. By employing Spatial Transformer Networks (STN), the model learns the displacement field that maps the source image to the target image. By warping the mask of the source image using the estimated displacement field, the mask of the target image can be obtained, serving as the segmentation result. The existing image registration models have shown

promising results in medical image analysis. However, very limited research works achieved good performance on handling large image deformations. To address this gap, a multi-scale image registration framework inspired by traditional registration algorithms was introduced in this chapter. The framework generates a displacement field at each scale, and the finest scale's displacement field is calculated by combining up-sampled displacement fields from coarser scales successively. Experiments were conducted on a challenging local 2D dataset and a public 3D dataset. The results demonstrated the effectiveness of this approach in improving registration performance for large image deformations while preserving good diffeomorphic properties. Additionally, the study showed that the use of a mask-guided term effectively enhanced the registration accuracy in the masked region.

The key contributions of this proposed image registration method are shown as follows: (1) a versatile trainable image registration framework, which robustly handles different levels of image deformations; (2) a new residual displacement field learning strategy that preserves good diffeomorphic properties; (3) a mask-guided loss term that enhances local alignment performance; (4)effective combination of the global NCC similarity loss and the mask-guided loss to achieve good image alignment in both global and local image regions.

In **chapter 6**, a novel semi-supervised learning model is developed by combining the image segmentation model and the image registration model proposed in chapter 5. This joint model uses the pseudo-labelling idea, similar to the method in chapter 4, to achieve iterative model improvement. Unlike the method in chapter 4, this joint model produces a soft pseudo-mask for each of the unannotated images, which is generated by both the segmentation and registration models. The soft pseudo-masks simulate the combination of multiple physicians' annotations to achieve more accurate segmentation outcomes. The two models are subsequently refined using the soft dice as a loss function. Experimental results on a 2D brain MRI dataset demonstrated that the joint model achieved significant performance improvements over both indi-

vidual models, especially when trained with very a small number of annotated data.

The key contributions of the joint training framework are listed as follows: (1) a novel integrated image segmentation and registration framework for semi-supervised medical image segmentation; (2) a pseudo-mask generator to generate soft masks for unannotated data, which has been proven to be effective in improving both models iteratively; (3) a novel method to generate the final segmentation result and an associated confidence map using the soft mask produced by the joint model.

In conclusion, this thesis presents a pipeline from interactive image annotation to fully automatic image segmentation. The pipeline involves two stages: firstly, employing the software developed in chapter 3 to efficiently annotate a subset of the acquired data. Next, one of the semi-supervised methods described in chapter 4 and chapter 6 can be applied to develop the automatic segmentation model. The method in chapter 4 is more suitable for the case of segmenting the targets without a common structure across images (e.g. tumours). The method in chapter 6 is applicable for segmenting objects that share a common anatomical structure (e.g. bones, organs).

7.2 Limitations and Future Works

This section discusses the limitations of the proposed methods and provides some ideas on potential future works.

7.2.1 Representative Data Selection for Annotation

Semi-supervised learning is based on a limited number of annotated images. Therefore, the model performance could be highly dependent on the variety of the annotated images. For instance, in chapter 4, the pseudo-masks generated by the pre-trained model can be influenced by the annotated images. If the annotated images do not adequately represent the data distribution of the ma-

jority of the population, it can lead to a sub-optimal pseudo-mask generation, hence limiting the ability of model to learn effectively.

Similarly, in the methods introduced in chapter 5 and chapter 6, the choice of the source image for the image registration model is also crucial, which can directly affect the segmentation results. An unrepresentative source image could result in a poorly warped image and mask. In chapter 6, the pre-trained segmentation model is also affected by the selected annotated images. However, in comparison to the method in chapter 4, it is more resilient to handle this problem by incorporating an unsupervised image registration model.

To enhance the representativeness of the annotated data, as a future work, an additional pre-processing step can be included in the proposed semi-supervised learning methods. For the pre-processing step, unsupervised clustering methods can be applied to group the data into meaningful sub-groups. By balanced sampling of these sub-groups, representative data samples can be obtained. This approach aims to train the model with more balanced and unbiased data, consequently reducing bias in the model prediction process. Following this approach, the semi-supervised model can then incorporate both the unannotated data and the selected representative samples for model training.

7.2.2 Theoretical Proof for Diffeomorphic Property in Image Registration Model

The experiments in chapter 5 shows the proposed multi-resolution image registration framework is able to preserve the diffeomorphic properties in certain extent. However, the theoretical proof of why the multi-scale residual displacement field learning can help achieving it was not provided. Therefore, in future research, it is desirable to study it further from a theoretical point of view. Furthermore, the study will also explore the potential extension of this approach to other registration models.

7.2.3 Ensemble Learning on Medical Image Classification Task

The effectiveness of the ensemble technique was demonstrated in chapter 4. As an experimental research work, an extension of the ensemble technique was implemented on an image classification task. In detail, an automatic cardiac MRI quality estimation framework was proposed using ensemble and transfer learning. In this proposed method, multiple pre-trained models were initialised and fine-tuned on 2-dimensional image patches sampled from the training data. In the model inference process, decisions from these models are aggregated to make a final prediction.

This framework was evaluated on CMRxMotion grand challenge (MICCAI 2022) dataset, which is small, multi-class, and imbalanced. Furthermore, the final trained model was also evaluated on an independent test set provided by the CMRxMotion organisers. Our proposed method achieved the classification accuracy of 72.5% and Cohen's Kappa of 0.6309. It was ranked the top 1 in the CMRxMotion grand challenge. More details can be found in Publication 4.

7.2.4 Generative Modelling to Improve Model Training

Semi-supervised learning is an effective solution, when the annotated data is limited but a large number of unannotated data is available. However, in certain cases of medical applications, the whole available dataset could be limited. In this case, generative modelling could be useful to synthesis new data samples from limited data. A preliminary research work was conducted on using a GAN-based generative method to enlarge the training dataset. Due to the complexity of generating effective segmentation labels using GAN models, the initial experiments were performed on a simpler regression problem where only numbers are used as the label of images.

A brief description is provided as follows. Brain age estimation based on

magnetic resonance imaging (MRI) is an active research area in early diagnosis of some neurodegenerative diseases (e.g. Alzheimer, Parkinson, Huntington, etc.) for elderly people or brain underdevelopment for the young group. Deep learning methods have achieved the state-of-the-art performance in many medical image analysis tasks, including brain age estimation. However, the performance and generalisability of the deep learning model are highly dependent on the quantity and quality of the training data set. Both collecting and annotating brain MRI data are extremely time consuming. In this work, to overcome the data scarcity problem, a GAN based image synthesis method is proposed. Different from the existing GAN-based methods, a task-guided branch (a regression model for age estimation) is integrated to the end of the generator in GAN. By adding a task guided loss to the conventional GAN loss, the learned low dimensional latent space and the synthesised images are more task specific. It helps to boost the performance of the down-stream task by combining the synthesised images and real images for model training. The proposed method was evaluated on a public brain MRI data set for age estimation. The proposed method outperformed (statistically significant) a deep convolutional neural network based regression model and the GAN-based image synthesis method without the task-guided branch. More importantly, it enables the identification of age-related brain regions in the image space. The paper can be found in Publication 3.

7.2.5 Geometry-aware Image Segmentation

Many deep learning-based image segmentation models like U-Net treat each pixel independently as a pixel-wise classification problem without preserving any geometric and topological information. The models could generate anatomically invalid results for medical image analysis. Hence, incorporating prior shape/geometry information into model learning is highly desirable in medical image segmentation.

By visualising the segmentation results produced by the proposed image registration method in chapter 5 and chapter 6, it is observed that the geometric property can be preserved using image registration method. This is due to that the diffeomorphic deformation is able to provide smoothed geometrical transformation without damaging the original object structure.

However, the results in chapter 6 indicate that the quality of the masks generated by the registration model is normally worse than those generated by the image segmentation model. Therefore, enhancing the performance of the registration model for segmentation tasks is a potential avenue for future research. Inspired by the pseudo-mask fusion method in chapter 6, one conceivable approach to improve the model results is by employing multiple source images as templates and combining their results to create the final mask. This method should be able to significantly improve the quality of the final mask compared to using a single random source image.

7.2.6 Quality Control

In the field of medical imaging, ensuring the quality of the produced segmentation masks by automatic segmentation models is of utmost importance due to the high accuracy requirements in clinical analysis. One popular method for assessing the quality of segmentation masks is by using confidence maps. The joint training framework proposed in chapter 6 is capable of providing such a confidence map for each output, which indicates the agreement of multiple masks generated from both the image segmentation and image registration models.

Through examining the correlation between a confidence value generated from the confidence map and a similarity score (e.g. Dice coefficient) derived from the predicted masks and the ground truths, it can be determined if the confidence map can effectively infer the quality of the predicted mask. In future work, a thorough quantitative analysis is needed to demonstrate the

feasibility of using the confidence map for quality control purpose.

Bibliography

[1] A. Murtiyoso and P. Grussenmeyer, "Point cloud segmentation and semantic annotation aided by gis data for heritage complexes," in *8th International Workshop 3D-ARCH" 3D Virtual Reconstruction and Visualization of Complex Architecture"*, vol. 42. Copernicus Publications, 2019, pp. 523–528.

[2] T. Pham, "Semantic road segmentation using deep learning," in *2020 Applying New Technology in Green Buildings (ATiGB)*. IEEE, 2021, pp. 45–48.

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

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[5] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

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

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[9] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[10] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[11] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka *et al.*, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.

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

[13] T. Chan and L. Vese, "An active contour model without edges," in *International Conference on Scale-Space Theories in Computer Vision*. Springer, 1999, pp. 141–151.

[14] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020.

[15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[16] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 289–293.

[17] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[18] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[19] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9252–9260.

[20] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 729–738.

[21] Z. Xu and M. Niethammer, "Deepatlas: Joint semi-supervised learning of image registration and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 420–429.

[22] R. Szeliski, *Computer vision: algorithms and applications*. Springer Nature, 2022.

[23] D. Levi, N. Garnett, E. Fetaya, and I. Herzlyia, "Stixelnet: A deep convolutional network for obstacle detection and road segmentation." in *BMVC*, vol. 1, no. 2, 2015, p. 4.

[24] M. Kucharczyk, G. J. Hay, S. Ghaffarian, and C. H. Hugenholtz, "Geographic object-based image analysis: a primer and future directions," *Remote Sensing*, vol. 12, no. 12, p. 2012, 2020.

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

[26] J. C. Bezdek, L. O. Hall, and L. P. Clarke, "Review of MR image segmentation techniques using pattern recognition," *Medical Physics*, vol. 20, no. 4, pp. 1033–1048, 1993.

[27] R. Frank, T. Grabowski, and H. Damasio, "Voxelvise percentage tissue segmentation of human brain magnetic resonance images," in *Abstracts, 25th Annual Meeting, Society for Neuro-Science*. Society for Neuroscience, 1995, p. 694.

[28] E. R. Hancock and J. Kittler, "Edge-labeling using dictionary-based relaxation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 2, pp. 165–181, 1990.

[29] R. Nock and F. Nielsen, "Statistical region merging," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1452–1458, 2004.

[30] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 48, no. 3, pp. 259–279, 1986.

[31] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on information theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[32] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, 2011.

[34] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[35] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[36] P. Kohli, M. P. Kumar, and P. H. Torr, "P3 & beyond: Solving energies with higher order cliques," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[37] V. Vineet, J. Warrell, P. Sturgess, and P. H. Torr, "Improved initialization and gaussian mixture pairwise terms for dense random fields with mean-field inference." in *BMVC*, 2012, pp. 1–11.

[38] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern recognition*, vol. 26, no. 9, pp. 1277–1294, 1993.

[39] V. Vezhnevets and V. Konouchine, "Growcut: Interactive multi-label nd image segmentation by cellular automata," in *proc. of Graphicon*, vol. 1, no. 4. Citeseer, 2005, pp. 150–156.

[40] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–308, 2009.

[41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[44] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images." *ECCV (5)*, vol. 7576, pp. 746–760, 2012.

[45] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 978–994, 2010.

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

[47] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[48] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[49] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361.

[50] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[51] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[53] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[54] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[55] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[56] D. Molho, J. Ding, Z. Li, H. Wen, W. Tang, Y. Wang, J. Venegas, W. Jin, R. Liu, R. Su *et al.*, "Deep learning in single-cell analysis," *arXiv preprint arXiv:2210.12385*, 2022.

[57] S. Kaymak, P. Esmaili, and A. Serener, "Deep learning for two-step classification of malignant pigmented skin lesions," in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE, 2018, pp. 1–6.

[58] Ş. Öztürk and U. Özkaya, "Skin lesion segmentation with improved convolutional neural network," *Journal of digital imaging*, vol. 33, pp. 958–970, 2020.

[59] N. Zhang, S. Francis, R. A. Malik, and X. Chen, "A spatially constrained deep convolutional neural network for nerve fiber segmentation in corneal confocal microscopic images using inaccurate annotations," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 456–460.

[60] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "Sa-unet: Spatial attention u-net for retinal vessel segmentation," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 1236–1242.

[61] Q. Yue, X. Luo, Q. Ye, L. Xu, and X. Zhuang, "Cardiac segmentation from lge mri using deep neural network incorporating shape and spatial priors," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 559–567.

[62] X. Sun, P. Garg, S. Plein, and R. J. van der Geest, "Saun: Stack attention u-net for left ventricle segmentation from cardiac cine magnetic resonance imaging," *Medical Physics*, vol. 48, no. 4, pp. 1750–1763, 2021.

[63] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 149–157.

[64] M. Jafari, S. Francis, J. M. Garibaldi, and X. Chen, "Lmisa: A lightweight multi-modality image segmentation network via domain adaptation using gradient magnitude and shape constraint," *Medical Image Analysis*, vol. 81, p. 102536, 2022.

[65] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of mr brain images with a convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[66] P. Kumar, P. Nagar, C. Arora, and A. Gupta, "U-segnet: fully convolutional neural network based automated brain tissue segmentation tool," in *2018 25th IEEE International conference on image processing (ICIP)*. IEEE, 2018, pp. 3503–3507.

[67] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[68] J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M. D. Hellmann, J. O. Deasy, G. Mageras, and H. Veeraraghavan, "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images," *IEEE transactions on medical imaging*, vol. 38, no. 1, pp. 134–144, 2018.

[69] M. El Adoui, S. A. Mahmoudi, M. A. Larhmam, and M. Benjelloun, "Mri breast tumor segmentation using different encoder and decoder cnn architectures," *Computers*, vol. 8, no. 3, p. 52, 2019.

[70] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[71] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

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

[73] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[74] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[75] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep convolutional neural network acoustic modeling," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4545–4549.

[76] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[77] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[78] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A gan-based image synthesis method for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105568, 2020.

[79] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[80] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[81] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[82] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, p. 180161, 2018.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Los Alamitos, CA, USA, June 2016, pp. 770–778.

[84] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.

[85] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[86] A. Lahiri, V. Jain, A. Mondal, and P. K. Biswas, "Retinal vessel segmentation under extreme low annotation: A generative adversarial network approach," *arXiv preprint arXiv:1809.01348*, 2018.

[87] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems*, 2015, pp. 3546–3554.

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

[89] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6728–6736.

[90] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer, 2019, pp. 554–565.

[91] W. Hang, W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin, "Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 562–571.

[92] K. Zheng, J. Xu, and J. Wei, "Double noise mean teacher self-ensembling model for semi-supervised tumor segmentation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1446–1450.

[93] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 820–833.

[94] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[95] Y. Zhou and S. Goldman, "Democratic co-learning," in *16th IEEE International Conference on Tools with Artificial Intelligence*. IEEE, 2004, pp. 594–602.

[96] L. Sun, J. Wu, X. Ding, Y. Huang, G. Wang, and Y. Yu, "A teacher-student framework for semi-supervised medical image segmentation from mixed supervision," *arXiv preprint arXiv:2010.12219*, 2020.

[97] D. Filipiak, A. Zapała, P. Tempczyk, A. Fensel, and M. Cygan, "Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding," *arXiv preprint arXiv:2211.03850*, 2022.

[98] Z. Feng, Q. Zhou, G. Cheng, X. Tan, J. Shi, and L. Ma, "Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum," *arXiv preprint arXiv:2004.08514*, vol. 1, no. 2, p. 5, 2020.

[99] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma, "Dmt: Dynamic mutual training for semi-supervised learning," *Pattern Recognition*, vol. 130, p. 108777, 2022.

[100] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 253–260.

[101] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[102] J.-C. Yoo and T. H. Han, "Fast normalized cross-correlation," *Circuits, systems and signal processing*, vol. 28, pp. 819–843, 2009.

[103] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2007, pp. 319–326.

[104] T. Lindeberg, "Scale invariant feature transform," 2012.

[105] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[106] S. S. M. Salehi, S. Khan, D. Erdogmus, and A. Gholipour, "Real-time deep pose estimation with geodesic loss for image-to-template rigid registration," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 470–481, 2018.

[107] Y. Sun, A. Moelker, W. J. Niessen, and T. van Walsum, "Towards robust ct-ultrasound registration using deep learning methods," in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1*. Springer, 2018, pp. 43–51.

[108] J. Zhang, "Inverse-consistent deep networks for unsupervised deformable image registration," *arXiv preprint arXiv:1809.03443*, 2018.

[109] T. C. Mok and A. C. Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 211–221.

[110] H. Li, Y. Fan, and A. D. N. Initiative, "Mdreg-net: Multi-resolution diffeomorphic image registration using fully convolutional networks with deep self-supervision," *Human Brain Mapping*, vol. 43, no. 7, pp. 2218–2231, 2022.

[111] Y. Lei, Y. Fu, J. Harms, T. Wang, W. J. Curran, T. Liu, K. Higgins, and X. Yang, "4d-ct deformable image registration using an unsupervised deep convolutional neural network," in *Artificial Intelligence in Radiation Therapy: First International Workshop, AIRT 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 1*. Springer, 2019, pp. 26–33.

[112] T. C. Mok and A. Chung, "Affine medical image registration with coarse-to-fine vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 835–20 844.

[113] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[114] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Joint learning of motion estimation and segmentation for cardiac mr image sequences," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 472–480.

[115] D. Mahapatra, Z. Ge, S. Sedai, and R. Chakravorty, "Joint registration and segmentation of xray images using generative adversarial networks," in *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*. Springer, 2018, pp. 73–80.

[116] M. S. Elmahdy, J. M. Wolterink, H. Sokooti, I. Išgum, and M. Staring, "Adversarial optimization for joint registration and segmentation in prostate ct radiotherapy," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 366–374.

[117] X. Chen, J. Graham, and C. Hutchinson, "Integrated frameworkfor simultaneous segmentation and registration of carpal bones," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 433–436.

[118] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.

[119] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1562–1573, 2018.

[120] P. Soille, "Generalized geodesy via geodesic time," *Pattern Recognition Letters*, vol. 15, no. 12, pp. 1235–1240, 1994.

[121] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016.

[122] K. Hoebel, V. Andrearczyk, A. Beers, J. Patel, K. Chang, A. Depeursinge, H. Müller, and J. Kalpathy-Cramer, "An exploration of uncertainty information for segmentation quality assessment." in *SPIE Medical Imaging*, 2020, p. 11313.

[123] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos - combined (ct-mr) healthy abdominal organ segmentation challenge data," Apr. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3362844

[124] X. Chen, J. Graham, C. Hutchinson, and L. Muir, "Automatic generation of statistical pose and shape models for articulated joints," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 372–383, 2013.

[125] ——, "Automatic inference and measurement of 3d carpal bone kinematics from single view fluoroscopic sequences," *IEEE transactions on medical imaging*, vol. 32, no. 2, pp. 317–328, 2012.

[126] D. Newitt, N. Hylton *et al.*, "Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials," 2016. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK

[127] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[128] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *International journal of biomedical imaging*, vol. 2013, 2013.

[129] P. Kohli, A. Osokin, and S. Jegelka, "A principled deep random field model for image segmentation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[130] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[131] A. E. Kavur, N. S. Gezer, M. Barış, Y. Şahin, S. Özkan, B. Baydar, U. Yüksel, Ç. Kılıkçıer, Ş. Olut, G. B. Akar *et al.*, "Comparison of semi-automatic and deep learning-based automatic methods for liver segmentation in living liver transplant donors," *Diagnostic and Interventional Radiology*, vol. 26, no. 1, p. 11, 2020.

[132] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients." *Psychological methods*, vol. 1, no. 1, p. 30, 1996.

[133] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[134] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[135] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Classification of urban lidar data using conditional random field and random forests," in *Joint Urban Remote Sensing Event 2013*. IEEE, 2013, pp. 139–142.

[136] J. Dolz, C. Desrosiers, L. Wang, J. Yuan, D. Shen, and I. B. Ayed, "Deep cnn ensembles and suggestive annotations for infant brain mri segmentation," *arXiv preprint arXiv:1712.05319*, 2017.

[137] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[138] J. Liu, L. Liu, B. Xu, X. Hou, B. Liu, X. Chen, L. Shen, and G. Qiu, "Bladder cancer multi-class segmentation in mri with pyramid-in-pyramid network," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 28–31.

[139] A. Hering, B. v. Ginneken, and S. Heldmann, "mlvirnet: Multilevel variational image registration network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 257–265.

[140] S. Zhao, Y. Dong, E. I. Chang, Y. Xu *et al.*, "Recursive cascaded networks for unsupervised medical image registration," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 600–10 610.

[141] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye, "Cyclemorph: cycle consistent unsupervised deformable image registration," *Medical Image Analysis*, vol. 71, p. 102036, 2021.

[142] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren, "Label-driven weakly-supervised learning for multimodal deformable image registration," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1070–1074.

[143] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1. Atlanta, Georgia, USA, 2013, p. 3.

[144] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-euclidean framework for statistics on diffeomorphisms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2006, pp. 924–931.

[145] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[146] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[147] A. Hoopes, M. Hoffmann, B. Fischl, J. Guttag, and A. V. Dalca, "Hypermorph: Amortized hyperparameter learning for image registration," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 3–17.

[148] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[149] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy," *Physics in Medicine & Biology*, vol. 50, no. 12, p. 2887, 2005.

[150] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[151] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk," *Frontiers in neuroinformatics*, vol. 7, p. 45, 2013.

[152] A. Yezzi, L. Zollei, and T. Kapur, "A variational framework for joint segmentation and registration," in *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*. IEEE, 2001, pp. 44–51.

[153] C. Chen, J. Graham, and C. Hutchinson, "Integrated frameworkfor simultaneous segmentation and registration of carpal bones," *18th IEEE International Conference on Image Processing*, pp. 433–436, 2011.

[154] K. M. Pohl, J. Fisher, W. E. L. Grimson, R. Kikinis, and W. M. Wells, "A bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, no. 1, pp. 228–239, 2006.

[155] D. Mahapatra and Y. Sun, "Joint registration and segmentation of dynamic cardiac perfusion images using mrfs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 493–501.

[156] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[157] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in ct," *European radiology*, vol. 29, pp. 1391–1399, 2019.

[158] J. Lourenço-Silva and A. L. Oliveira, "Using soft labels to model uncertainty in medical image segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 585–596.

Appendix A

List of Abbreviations

2D	Two Dimensional
3D	Three Dimensional
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long-Short Term Memory
GAN	Generative Adversarial network
FC	Fully Connection
FCN	Fully Convolutional Network
SVM	Support Vector Machine
CRF	Conditional Random Fields
DSC	Dice coefficient
NCC	Normalised Cross Correlation
GNCC	Global Normalised Cross Correlation
LNCC	Local Normalised Cross Correlation
MAE	Mean Absolute Error
MSE	Mean Squared Recall