Detection of Cervical Spine Vertebrae on MRI towards Improving Multi-modal Image Registration

by

Jonathan Chu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2024

© Jonathan Chu 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Jonathan Chu was the sole author of Chapters 1, 2, and 5, which were written under the supervision of Dr. Stewart Mclachlin, Dr. Alexander Wong, and Dr. Michael Hardisty and were not written for publication. The thesis was solely written by Jonathan Chu, with intellectual input contributed by all supervisors.

Dr. Stewart McLachlin and Dr. Michael Hardisty were the primary co-investigators on the Canadian Institutes of Health Research (CIHR) grant which supported this work.

Research presented in Chapter 3: The research was conducted at the University of Waterloo by Jonathan Chu, under the supervision of Dr. Stewart McLachlin, Dr. Alexander Wong, and Dr. Michael Hardisty of Sunnybrook Research Institute. Dr. Stewart McLachlin, Dr. Michael Hardisty, and Jonathan Chu contributed to the study design, with Dr. Alexander Wong supporting the work by providing feedback. Jonathan Chu was the primary contributor in generating a novel annotated dataset, producing code, and analyzing results, in addition to writing the chapter.

Research presented in Chapter 4: This research was conducted at the University of Waterloo by Jonathan Chu, under the supervision of Dr. Stewart McLachlin, Dr. Alexander Wong, and Dr. Michael Hardisty of Sunnybrook Research Institute. Dr. Stewart McLachlin, Dr. Michael Hardisty, Raj Ranabhat and Jonathan Chu contributed to the study design, with Dr. Alexander Wong supporting the work by providing feedback. Jonathan Chu and Raj Ranabhat contributed in producing code for multi-sequence and multi-modal deformable registration pipelines. Jonathan Chu was the primary contributor in the analysis of results, in addition to writing the chapter.

Abstract

Multi-modal image registration, such as MRI to CT, is an important but often challenging aspect for clinical image analysis. It has applications in treatment planning requiring image fusion, or inter-subject, atlas-based analyses, as well as longitudinal analyses. Multi-modal image registration in the cervical spine presents extra challenges because of the variability in the field of view (FoV) of magnetic resonance imaging (MRI) of the spinal column between different image series, with cervical vertebrae having a similar appearance leading to many local registration minima.

This thesis explores methods to detect, localize, and label cervical vertebrae and the spinal cord in anatomical MRI, focusing on the application of deep learning techniques. Specifically, we generated a custom annotated dataset of the cervical spine MRI, based on the Spine Generic dataset [1], resulting in 149 T1w and 100 T2w labelled images. We then successfully trained a Mask R-CNN model [2] and utilized a weighted directed acyclic graph (DAG) to leverage the sequential hierarchy of the vertebrae to filter detections for the cervical vertebrae and spinal cord detection task. This resulted in state-of-the-art performance, where the model was robust to varying cervical spine FoVs. Lastly, we integrated the detector model into a multi-sequence and multi-modal deformable image registration pipeline, where the inference results were used to crop images to an appropriate FoV and seed initial alignment, prior to deformable registration. The multi-sequence pipeline utilized the generated custom dataset and successfully demonstrated the use of a trained cervical vertebrae detector for FoV cropping prior to affine and deformable registration. The multi-modal pipeline was created, adopting from the multi-sequence pipeline, but with an additional CT vertebrae detection branch and utilized affine registration prior to FoV cropping. It was evaluated utilizing a prospectively maintained clinical dataset (SpineMets) provided by collaborators at Sunnybrook Research Institute, containing treatment planning CT and MRI scans. The pipeline demonstrated the ability to successfully perform deformable registration on multi-modal imaging. However, performance on this clinical dataset was limited by the Mask R-CNN model.

The methodology developed in this thesis demonstrates the potential of deep learning object detection models combined with leveraging vertebrae hierarchy to provide robust detections of vertebrae in the cervical spine. In addition, we illustrated the application of these models to determine vertebrae FoVs towards improving multi-modal deformable image registration. By leveraging these techniques, this research can be applied to automate and increase efficiency of cervical spine image registration for clinical needs, resulting in reduced cost and burden on clinical experts.

Acknowledgements

I would like to thank my supervisors Dr. Stewart McLachlin and Dr. Alexander Wong for their guidance and support throughout my MASc program. I would also like to extend my thanks to the Centre for Bioengineering and Biotechnology and the NSERC CREATE program for their support in funding the project.

I would also like to thank Dr. Michael Hardisty from Sunnybrook Research Institute for his insights and mentorship. In addition, I would like to thank Raj Ranabhat for his support and contributions to the projects.

Lastly, I am grateful for my friends that have supported me throughout my undergraduate and graduate experiences, shaping who I am today.

Dedication

This is dedicated to my family who have supported me and my goals since day one.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures x

List of Tables xvii

1 Introduction 1

1.1 Clinical Applications of Spine Imaging 1

1.2 Automated Spine Registration Challenges 2

1.3 Thesis Objectives and Contributions 3

2 Background 4

2.1 Spine Imaging 4

2.1.1 Anatomy of the Spine 4

2.1.2 Magnetic Resonance Imaging (MRI) 7

2.2 Vertebral Detection . . . 10
2.2.1 Object Detection and Deep Learning . . . 10
2.3 Multi-Modal Image Registration . . . 12
2.3.1 Image Registration Methodology . . . 13
2.3.2 Affine Image Registration . . . 14
2.3.3 Deformable Image Registration . . . 14
2.4 Translation to Clinical Applications . . . 15

3 Vertebral Detection and Labelling using Deep Learning for Spine MRI Registration 16
3.1 Introduction . . . 16
3.1.1 Related Works . . . 17
3.2 Methods . . . 20
3.2.1 Dataset Generation . . . 20
3.2.2 Mask R-CNN Vertebra Detector Model . . . 24
3.2.3 Evaluation Criteria . . . 29
3.3 Results and Discussion . . . 33
3.3.1 Qualitative Evaluation of Mask R-CNN Model . . . 34
3.3.2 Quantitative Evaluation of Mask R-CNN Model . . . 34
3.3.3 Ablation Experiments . . . 39
3.3.4 State-of-the-Art Detection Model Comparison . . . 41
3.4 Conclusions . . . 44

4 Application of Vertebral Detection Models for Multi-Modal Medical Deformable Image Registration 45
4.1 Introduction . . . 45
4.2 Methods . . . 46
4.2.1 Multi-Sequence Deformable Registration Pipeline . . . 46
4.2.2 Pipeline Evaluation Criteria . . . 51

4.3 Results and Discussion . . . 52
4.3.1 Qualitative Evaluation of the Multi-Sequence Deformable Image Registration Pipeline . . . 53
4.3.2 Quantitative Evaluation of the Multi-Sequence Deformable Image Registration Pipeline . . . 53
4.4 Application to Clinical Data . . . 55
4.4.1 Spine Metastasis Dataset . . . 55
4.4.2 Multi-Modal Deformable Image Registration Pipeline for Clinical Dataset . . . 57
4.4.3 Vertebrae and Spinal Cord Detection . . . 57
4.4.4 Deformable Multi-Modal Registration . . . 63
4.5 Conclusions . . . 68

5 Thesis Conclusion **69**
5.1 Significance of Research Contributions . . . 69
5.2 Limitations and Future Work . . . 69
5.2.1 Limitations and Future Work for the Mask R-CNN Cervical Vertebrae and Spinal Cord Detector . . . 70
5.2.2 Limitations and Future Work for the Multi-Modal Deformable Image Registration Pipeline . . . 71

References **72**

APPENDICES **80**

A Mask R-CNN Detection Model Training and Validation Curves **81**

B SpineMets MRI Inference Results **84**
B.1 Mid-Sagittal Slice of MRI Volume Input . . . 84
B.2 Manually Selected Mid-Sagittal Slice of MRI Volume Input . . . 86

C SpineMets T2w Affine and Deformable Registration Results **88**

List of Figures

2.1 Anatomy of the human spine, or vertebral column. It consists of three distinct regions: (1) the cervical spine, (2) the thoracic spine, and (3) the lumbar spine (Wikimedia Commons [3]) 5

2.2 Anatomy of a Spinal Vertebra. (Wikimedia Commons [4]) 6

2.3 Unique features of C1 and C2 vertebrae to allow for movement of the neck. a) C1 vertebrae anatomy, b) C2 vertebrae anatomy (Wikimedia Commons [5, 6]) 6

3.1 A single subject T1w and T2w MRI from the spine generic dataset [1]. a) T1w image containing entire skull, cervical spine, extending into the thoracic spine. The image has an isotropic voxel size of 1 mm^3 and has a voxel resolution of approximately 192x260x320 (RAS). b) T2w image containing the cervical spine, extending into the thoracic spine. The image has an isotropic voxel size of 0.8 mm^3 with a voxel resolution of approximately 64x320x320 (RAS) 21

3.2 Dataset generation workflow used to obtain ground truth annotations of the T1w Spine Generic Dataset. A similar process is used to generate the T2w dataset. a) The raw Spine Generic T1w MRI images of healthy subjects. b) Initial segmentation and labelling of the spinal cord using SCT Deepseg with different colours representing different instance segmentations of the vertebral levels on the spinal cord. A semantic segmentation of the whole spinal cord is also generated. c) Registration of PAM50 template vertebral column using SCT. d) Relabelling registered PAM50 template into individual vertebra and incorporation of spinal cord segmentation in 3D Slicer. e) Cropping to the cervical spine (C1-C7) after quality checks. f) Manual segmentation clean-up and removal of any thoracic vertebra segmentations and labels visible highlighted by red arrows. g) Generate various FoV windows of the cervical spine with a minimum of 3 vertebra. 22

3.3 Mask R-CNN architecture for object detection and instance segmentation. Feature maps are extracted from the input image using a ResNet backbone FPN, with object RoIs proposed by an RPN. A RoI Align process then extracts the regions from the feature maps, followed by the classification, bounding box regression, and instance segmentation (mask) branches for their respective tasks. 25

3.4 General representation of the post-inference graph structure. Each node consists of the predicted label of a vertebra detection, n, and edge weight, w, representing the model confidence of the detection. N represents the vertebrae label and k represents the k^{th} prediction. Each row represents a vertebral level, or additional generated neighbour nodes with different vertebral level labels, with each column representing a prediction from the model. The the nodes from the same prediction, k, cannot connect as they are either the model prediction, or neighbour node with the same mask and bounding box values. The T and B nodes correspond to the top and bottom of the detections and have weights of zero. The nodes of the weighted DAG are connected based on three main rules. The red path is an example of the longest path from T to B when solving the weighted DAG. 30

3.5 Example of the graph filtering four inference predictions with ground truth vertebrae and anatomical location of each prediction outlined by the dashed box from top to bottom. Orange nodes are generated neighbouring nodes due to the prediction having a confidence score less than 90%. The longest path is then calculated to filter and relabel predictions, highlighted in red. 31

3.6 Example inference outputs of the T1w Mask R-CNN model for cervical vertebra and spinal cord detection show excellent vertebra detection performance with varying vertebra FoV. Each inference output is from a different subject. a) Detection of C1-C7. b) Detection of C4-C6. c) Detection of C1-C6 with C7 extra detection. 34

3.7 Example inference outputs of the T2w Mask R-CNN model for cervical vertebra and spinal cord detection show excellent vertebra detection performance with varying vertebra FoV. Each inference output is from a different subject. a) Detection of C1-C7. b) Detection of C3-C6. c) Detection of C4-C6 with C7 extra detection. 35

3.8 Normalized confusion matrices of the (a) T1w and (b) T2w models during test set inference. Class 0 is the background class, showing undetected vertebrae (false negatives) and extra vertebrae detections (false positives). 38

3.9 Inference results of T1w whole vertebrae detection vs. vertebral body only detection of the same subject. a) T1w whole vertebrae detection of C1-C5, with extra detection of C6. b) T1w vertebral body detection of C1-C5, without extra detection. 41

4.1 Intra-subject, multi-sequence image registration pipeline for T1w to T2w deformable registration. a) Input mid-sagittal slice. b) Feature detection using Mask R-CNN cervical vertebrae and spinal cord MRI detection model. c) Cropping fixed or moving image to limiting FoV. d) Initial affine registration using bounding box centroids of moving image overlaid on fixed image. e) Continuous optimization-based deformable image registration of moving image overlaid on fixed image. 47

4.2 Regeneration of spinal cord bounding box due to cropping in the SI axis. a) Cropped image (originally C1-C7) in SI axis results in an oversized spinal cord bounding box in the AP axis. b) Regenerated spinal cord bounding box based on the cropped spinal cord segmentation results in a shrunken bounding box in AP axis. 49

4.3 Moving image orientation potentially mirroring and rotating when obtaining affine transformation matrix. a) Cropped fixed and moving images with vertebrae centroids. b) Mirrored and rotated moving image from applied affine transformation without spinal cord centroid. c) Affine transformation with vertebrae and spinal cord centroids enforce image orientation. 50

4.4 SCT segmentation of the vertebral level on the spinal cord used for quantitative evaluation of the multi-sequence deformable registration pipeline. a) T1w moving image. b) T2w cropped fixed image. 52

4.5 Spine Generic T1w to T2w deformable registration results of two subjects. The moving T1w image is superimposed on the cropped fixed T2w image and T2w inference vertebrae segmentation. a) Moving T1w image and fixed T2w image in original physical space. b) Affine aligned moving T1w to T2w image using inference bounding box centroids. c) Deformable registration applied to moving T1w image. 54

4.6 A single subject from the SpineMets clinical dataset collected by Sunnybrook Research Institute. Pathology observed in the C4 vertebrae. a) CT scan containing C1-T3 vertebrae. b) T1w scan containing C2-C7. c) T2w scan containing C2-C7. 56

4.7 Multi-modal spine image registration pipeline for CT to MRI (T1w or T2w) deformable registration. This pipeline adopts from the multi-sequence spine image registration pipeline, with the addition of the VertDetect head for CT vertebrae detection and modifications for the affine registration and FoV cropping algorithm. 58

4.8 Example inference outputs of T1w MRI scans from the SpineMets clinical dataset using the mid-sagittal slice of the volume. a) Subject 1 contains C1-C4 with inference results incorrectly detecting C6. b) Subject 2 contains C2-C7 with inference results detecting C3-C7 and the spinal cord correctly. c) Subject 3 contains C2-C6 with no detections in inference. d) Subject 4 contains C1-C5 with only C2 detected correctly. 60

4.9 Example inference outputs of T2w MRI scans from the SpineMets clinical dataset using the mid-sagittal slice of the volume. a) Subject 1 contains C1-C5 with inference results detecting C2-C5 and the spinal cord correctly. b) Subject 2 contains C2-C7 with inference results detecting C2-C7 and the spinal cord correctly. c) Subject 3 contains C1-C6 with inference results detecting C1-C4 and the spinal cord correctly, with a mislabelled detection of C6 as C5. d) Subject 4 contains C1-C5 with C1-C4 and the spinal cord correctly detected. 61

4.10 Example VertDetect inference outputs of CT scans from the SpineMets clinical dataset. a) Input CT scan on Subject 1 containing C1-T4. b) Inference results containing labelled 3D segmentations of vertebrae. 62

4.11 Inference outputs of T1w MRI scans from the SpineMets clinical dataset using the manually selected mid-sagittal slice. a) Subject 1 contains C1-C4 with inference results correctly detecting C2-C3 and the spinal cord. b) Subject 2 contains C2-C7 with inference results detecting C3-C7 and the spinal cord correctly. Note that the mid-sagittal slice of the volume was the ideal slice for this subject. c) Subject 3 contains C1-C6 with C6 correctly detected, and C3 incorrectly detected as C5. d) Subject 4 contains C1-C5 with an incorrect detection of C6 on the C5 posterior element. 63

4.12 Inference outputs of T2w MRI scans from the SpineMets clinical dataset using the manually selected mid-sagittal slice. a) Subject 1 contains C1-C5 with inference results correctly detecting C1-C5 and the spinal cord. b) Subject 2 contains C2-C7 with inference results detecting C2-C7 and the spinal cord correctly, with a incorrect detection of C1 in the air gap. c) Subject 3 contains C1-C6 with inference results detecting C1-C4 and the spinal cord correctly, with a mislabelled detection of C6 as C5. Subject 4 contains C1-C5 with C1-C4 and the spinal cord correctly detected 64

4.13 Multi-modal registration of Subject 1 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w MRI volume (C2-C5 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume. Moving T2W MRI results in vertebrae better fitting the fixed CT vertebrae segmentations. . . 65

4.14 Multi-modal registration of Subject 4 with T2w volume and VertDetect vertebrae detections on CT volume. One undetected vertebrae from inference. a) Affine registration of T2w volume (C1-C4 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w Volume to CT volume. Moving T2W MRI results in vertebrae better fitting the fixed CT vertebrae segmentations even with the undetected vertebrae from inference. 66

4.15 Multi-modal registration of Subject 3 with T2w volume and VertDetect vertebrae detections on CT volume. One undetected and mislabelled vertebrae from inference. a) Affine registration of T2w volume (C1-C4 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w Volume to CT volume. Moving T2W MRI results in vertebrae better fitting the fixed CT vertebrae segmentations even with the undetected and mislabelled vertebrae from inference. 67

A.1 T1w whole vertebrae model training and validation loss curves. 82

A.2 T2w whole vertebrae model training and validation loss curves. 83

B.1 Inference outputs of T1w MRI scans SpineMets clinical dataset using the mid-sagittal slice of the volume of remaining subjects. a) Subject 5 contains C1-C6 with no detections. b) Subject 6 contains C1-C5 with no detections. c) Subject 7 contains C1-C5 with no detections. 84

B.2 Inference outputs of T2w MRI scans SpineMets clinical dataset using the mid-sagittal slice of the volume of remaining subjects. a) Subject 5 contains C2-C6 with inference results correctly detecting C2-C5 and the spinal cord, with C6 undetected. b) Subject 6 contains C2-C5 with detections of C3-C5 and the spinal cord. c) Subject 7 contains C1-C5 with C3-C4 and the spinal cord correctly detected. C2 is correctly detected, but the segmentation and bounding box extends past the vertebrae into soft tissue in the superior direction. 85

B.3 of T1w MRI scans from the SpineMets clinical dataset using the manually selected mid-sagittal slice. a) Subject 5 contains C1-C6 with no detections. b) Subject 6 contains C1-C5 with no detections. d) Subject 7 contains C1-C5 with no detections. 86

B.4 Inference outputs of T2w MRI scans SpineMets clinical dataset using the manually selected mid-sagittal slice of remaining subjects. a) Subject 5 contains C2-C6 with inference results correctly detecting C2-C5 and the spinal cord, with C6 undetected. b) Subject 6 contains C2-C5 with detections of C2-C5 and the spinal cord. c) Subject 7 contains C1-C5 with C3-C4 and the spinal cord correctly detected. C2 is correctly detected, but the segmentation and bounding box extends past the vertebrae into soft tissue in the superior direction. 87

C.1 Multi-modal registration of Subject 2 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C7 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume. 88

C.2 Multi-modal registration of Subject 5 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C5 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume. 89

C.3 Multi-modal registration of Subject 6 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C5 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume. 89

C.4 Multi-modal registration of Subject 7 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C4 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume. 90

List of Tables

3.1 Quantitative metrics used to evaluate the Mask R-CNN model demonstrate robustness in detecting cervical whole vertebra. 36

3.2 Average Precision (AP) values of each vertebra and spinal cord from object detection averaged across IoU thresholds between 0.50 and 0.95 at 0.05 intervals. 36

3.3 Ablation experiment verifying the effectiveness of the graph filter. 39

3.4 Ablation experiment verifying the effectiveness of training to detect the whole vertebrae rather than only the vertebral body. 40

3.5 Ablation experiments to determine score threshold effect on graph filtered inference results using T1w whole vertebrae model. 42

3.6 Ablation experiments to determine score threshold effect on graph filtered inference results using T2w whole vertebrae model. 42

3.7 Comparison of cervical spine Mask R-CNN vertebrae and spinal cord detector to Faster R-CNN detection results for mixed T1w/T2w scans of lumbar/thoracic vertebrae. 43

3.8 Comparison of cervical spine MRI Mask R-CNN vertebrae and spinal cord detector to state-of-the-art deep learning vertebrae MRI detection models. 44

4.1 Quantitative Dice coefficient metric used to evaluate the deformable registration performance from the multi-sequence spine image registration pipeline. 53

Chapter 1

Introduction

Medical imaging, including magnetic resonance imaging (MRI) and computed tomography (CT), is vital for many clinical workflows. For example, the use of MRI and/or CT is a gold standard in the diagnosis and treatment of complex spinal disorders, ranging from degeneration to trauma and cancer care. For accurate clinical decision-making, multi-modal medical imaging is often needed, with image registration utilized for anatomical alignment to assist with clinical planning and evaluation. However, automated image registration is a difficult problem due to the differing scanning parameters and spatial field of views (FOVs) in images. This chapter outlines the applications of spine imaging in clinical workflows, registration techniques of spine imaging, and roadblocks preventing automated spine registration. This chapter also includes an outline of the thesis objectives and contributions.

1.1 Clinical Applications of Spine Imaging

Medical imaging of the spine is a standard of care that can be used for pre-operative diagnosis and treatment planning, to post-operative evaluation and follow-up of treatment progression. Specifically, spine MRI is commonly used for the detection of pathologies such as infections, metastases, and disc abnormalities [7]. In comparison, spine CT provides an excellent visualization of the vertebrae structures for detection of bony abnormalities [8] and landmarking [9].

When used in clinical planning, spine imaging can influence the treatment method of choice [10], to the planning of the surgery navigation procedures [11]. Another common

use of spine MRI and CT in planning and clinical decision making is image segmentation and multi-modal image registration, or the spatial alignment of images. For example, in radiation oncology, CT scans are a standard of care used to determine radiation dosage for radiation therapy. Additional MRI scans can then be performed, with subsequent image registration of the CT and MRI scans. This is due to the different anatomy highlighted in the patient by the two imaging modalities; soft tissue in MRI, and bony structures in CT. Planning of radiation therapy includes determining the target, which is located under MRI, while landmarking is done based on the relative bony anatomy surrounding the cancerous tissue [12]. Inter-subject image registration can also be used to compare subject anatomy to an atlas or reference. However, registration of spine imaging is notoriously difficult due to the requirement of time consuming manual landmarking and segmentation, leading to the need of improved clinical tools.

1.2 Automated Spine Registration Challenges

Registration of spinal anatomy from different imaging can provide critical information for planning clinical procedures, but is often gated by time consuming manual landmarking and segmentation requiring experts. In addition, the vertebrae being annotated can be very similar in size and shape and contain differing FoVs, further complicating the registration process. To streamline the process, the development of an automated spine registration pipelines is required, yet there is a gap in the current literature of end-to-end automated spine registration pipelines. Additionally, there are recognized challenges for the implementation, including the requirement of both images containing similar FoVs of the anatomical and spatial features used for registering between two or more images. The registration process can be assisted through an initial alignment step of the important anatomical features.

Initial alignment of spine images is straightforward when the images each contain the same number of vertebral levels. However, due to patient positioning, scanning targets, and time constraints, resultant clinical images may contain different spine FoVs with varying vertebral levels and number of vertebrae. For example, a CT may contain the entire the entire cervical spine (C1-C7), whereas an MRI of the same subject may only contain a smaller section of interest (C3-C6). Generally, MRIs have a smaller FoV due to the time and cost associated with the imaging modality. This mismatch in FoVs and differing imaging characteristics often leads to the requirement of manual, or semi-automated, registration where one must manually detect vertebra present and crop the images to the same FoV. This can be especially difficult when common anatomical landmark, such as the unique C2

vertebra, are not present to assist in classifying other vertebra that have similar appearance. This gap in automated detection of vertebral levels for multi-modal image registration remains an outstanding challenge needed in spine imaging.

1.3 Thesis Objectives and Contributions

The goal of this thesis was to develop and evaluate a new approach for the detection of cervical vertebrae on MRI towards improving multi-modal image registration. This challenge necessitates dealing with multi-modal spine imaging with varying FoV. Proposed contributions of this thesis include: 1) development of a deep-learning based algorithm for cervical vertebrae and spinal cord detection in MRI, 2) assessment of the detection algorithm on different spine imaging datasets with varying FoVs, and 3) integration of the detection algorithm within a multi-modal spine image registration pipeline.

This thesis is organized as follows:

- Chapter 2 introduces background information on spine imaging, methods of vertebral detection, and multi-modal medical image registration techniques.
- Chapter 3 details the development of a unimodal deep-learning based approach for cervical vertebrae and spinal cord detection and labelling.
- Chapter 4 outlines the utilization of the vertebrae detection algorithm in a multi-sequence and multi-modal deformable image registration pipeline on clinical spine data.
- Chapter 5 concludes and summarizes the results in the wider context of spinal vertebrae detection and image registration. Challenges towards clinical adoption and future works are also discussed.

Chapter 2

Background

This chapter provides an overview of essential background information relevant to the thesis. It encompasses sections on spine imaging, vertebral detection, and multi-modal image registration.

2.1 Spine Imaging

Medical imaging of the spine is a critical tool in many clinical workflows, where it is used in diagnosing and treating various spinal conditions ranging from trauma to cancerous tumors. Different imaging modalities allow for the capture of detailed images to visualize the spine's structure and surrounding anatomy, allowing for clinicians to accurately assess, plan, and provide therapy. The main types of spine medical imaging include x-ray, computed tomography (CT), and magnetic resonance imaging (MRI) [13, 14]. This thesis is focused on cervical spine MRI.

2.1.1 Anatomy of the Spine

The human spine, or vertebral column, is a complex structure that plays a crucial role in supporting the body and protecting the spinal cord. The spine, as seen in Figure 2.1 typically comprises 24 individual vertebrae, in addition to sacrum and coccyx at the base of the spine [15]. The 24 individual vertebra can be split into three distinct regions: (1) cervical spine, (2) thoracic spine, and (3) lumbar spine. The vertebrae are numerically labelled from top to bottom based on the vertebrae region. The cervical spine is the

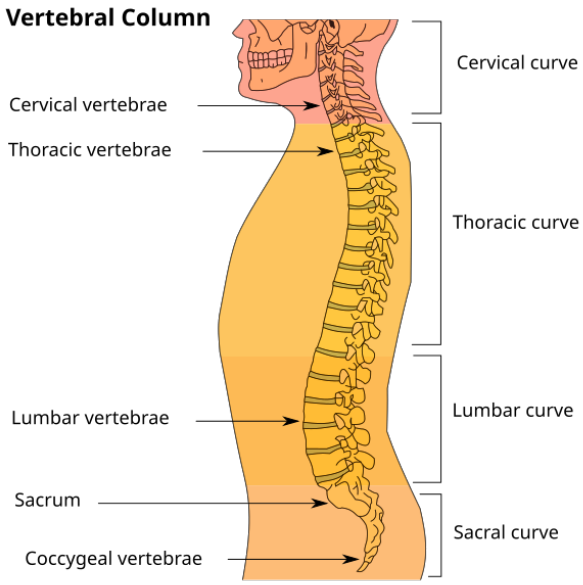


Figure 2.1: Anatomy of the human spine, or vertebral column. It consists of three distinct regions: (1) the cervical spine, (2) the thoracic spine, and (3) the lumbar spine (Wikimedia Commons [3])

uppermost part of the spine, consisting of seven vertebrae labelled from C1-C7. The thoracic spine consists of the mid-back section of the spine and consists of twelve vertebrae labelled from T1-T12. Lastly, the lower-back region is called the lumbar spine, consisting of five vertebrae labelled from L1-L5.

Vertebrae

Vertebrae are the individual bones that stack to form the spinal column. As seen in Figure 2.2, each vertebra consists of a vertebral body in the anterior region, a central hole that forms the spinal canal protecting the spinal cord, and the posterior bony elements like the spinous process

The cervical vertebrae, located in the neck, are the smallest and are the most mobile [15, 16]. To allow for neck movements such as nodding up and down and rotating the head from side to side, the C1 (Atlas) and C2 (Axis) vertebra of the cervical spine are quite distinct as seen in Figure 2.3 a) Figure 2.3 b), respectively. The C1 vertebra is ring shaped and supports the skull, whereas the C2 vertebra has a unique structure called the odontoid

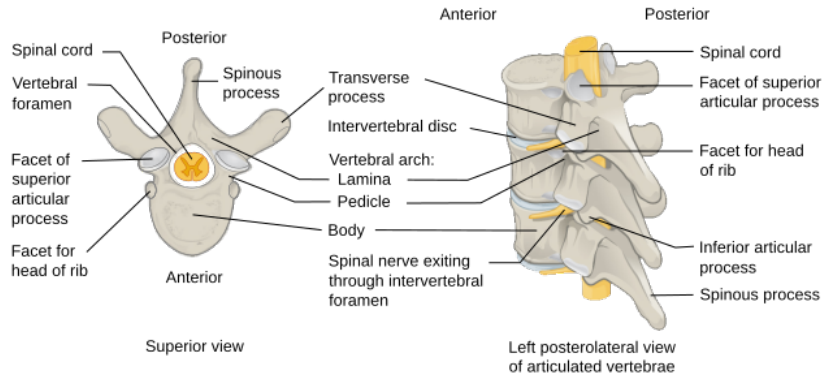


Figure 2.2: Anatomy of a Spinal Vertebra. (Wikimedia Commons [4])

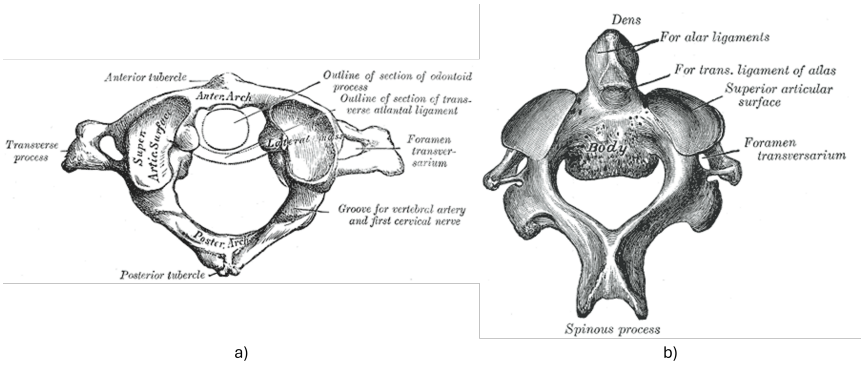


Figure 2.3: Unique features of C1 and C2 vertebrae to allow for movement of the neck. a) C1 vertebrae anatomy, b) C2 vertebrae anatomy (Wikimedia Commons [5, 6])

process, or dens, in which the C1 vertebrae sits around to allow for the turning motion of the head. Vertebrae of the sub-axial region of the cervical spine (C3-C7) all have a similar appearance, but with posterior elements increasing in size. The thoracic vertebrae are larger, connected to the ribs, and less mobile than the cervical vertebrae. The lumbar vertebrae are the largest vertebrae as they support the weight of the upper body.

Intervertebral Discs

Each of the individual vertebra are separated by intervertebral discs (Figure 2.2) that are sandwiched between the vertebral bodies, with the exception of the first and second cervical

vertebrae [15]. The discs act as a cushion to absorb shock and allow flexibility of the spine. Each disc has a soft, gel-like center called the nucleus pulposus, and is surrounded by a tough outer run called the annulus fibrosus [17]. The discs allow for the flexibility and movement of the spine.

Facet Joints

Facet joints (Figure 2.2) are small joints located between the vertebrae and behind adjacent vertebrae along the posterior elements. They provide stability to the spine and allow for smooth movement. The joints are lined with cartilage, which reduces friction and allows for flexibility.

Spinal Cord

The spinal cord is contained within the central canal of the vertebrae and is protected by the surrounding bony structure as seen in Figure 2.2. The spinal cord is a vital part of the central nervous system that transmits signals between the brain and the rest of the body. Nerve roots branch out from the spinal cord though openings in the vertebrae to control motor and sensory functions of the body.

2.1.2 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) is a 3D non-invasive medical imaging technique, or modality, that utilizes strong magnetic fields and radio-frequency (RF) pulses to measure the response of hydrogen nuclei (protons) in intracellular fluids [13]. The MRI machine generates and applies a strong magnetic field to first align the protons in the body's hydrogen atoms. RF pulses are then applied to disturb the alignment of those protons. As the protons return to their original, steady-state alignment under the magnetic field, they emit signals detected by the MRI machine. Then, the MRI machine utilizes the time delta, or relaxation time, of the protons shifting back to steady-state alignment to distinguish between tissues, with water content of the tissue determining signal intensity.

This process results in a detailed, multi-planar image with excellent anatomical and spatial resolution. MRI is particularly effective in imaging soft tissues like the brain, spinal cord, muscles and ligaments, while still being able to image bony structures. Additionally, since MRI does not use harmful ionizing radiation, it is the diagnostic procedure of choice for most spinal pathologies and routinely ordered in a clinical setting [14].

In anatomical MRI, there are two primary imaging sequences: T1-weighted (T1w) and T2-weighted (T2w) imaging. T1w imaging provides high-resolution images where fat appears bright and water appears dark. This is useful for visualizing anatomy and detecting fat containing structures such as white matter in the brain and bone marrow. In comparison, T2w imaging highlights anatomy based on water content, where high water content appears bright, and low water content appears dark. This is useful for detecting anatomy with fluids such as the cerebral spinal fluid along the spinal cord, or detecting pathology such as tumours or edemas that have higher water content compared to surrounding tissue. In addition, scans utilize an anatomical coordinate system that is described by the standard anatomical position of a human. This includes the axial plane, coronal plane, and sagittal plane. Often this is with respect to RAS (right, anterior, sagittal). The image visualizations then utilize an image coordinate system (ijk), describing how the image was captured with respect to the anatomy.

Clinical Relevance of MRI in Spine Imaging

Images captured using MRI have superior soft tissue contrast, leading MRI to become an indispensable tool for the imaging, evaluation, and management of various spinal conditions. In addition, MRI is a non-invasive and non-ionizing procedure, making it a standard of care in clinical settings. When used in spine imaging, MRI can be used to observe spinal degeneration and trauma, spinal cord injury, as well as bone metastasis and other cancers.

MRI is highly sensitive in detecting acute soft tissue and spinal cord injuries, making it crucial in the assessment of spinal trauma. It is particularly effective in identifying ligamentous injuries, disc herniations, and epidural hematomas, which are often missed by other imaging modalities like CT scans [18]. Additionally, MRI is the gold standard in evaluating the extent of damage to the spinal cord and surrounding structures. This is due to the detailed images capturing intramedullary abnormalities such as an edema and hemorrhage, which are critical for determining the severity of the injury and guiding treatment decisions [19]. Early MRI can also help predict long-term neurological outcomes by assessing the degree of spinal cord compression and the presence of any ligamentous instability [19] and help plan appropriate surgical and conservative treatments[18].

MRI also plays a critical role in the detection and characterization of metastasis and primary spinal tumors. Its high contrast resolution allows for the differentiation between benign and malignant lesions, assessment of the extent of bone marrow involvement, and evaluation of any associated spinal cord compression. This information is vital for tailoring treatment strategies such as surgery, radiation, or chemotherapy [10].

Challenges of MRI in Spine Imaging

While MRI is a powerful and widely adopted tool for spine imaging, it does face several challenges that can impact the effectiveness and accuracy of the images. These challenges in spine MRI imaging can be broadly categorized into issues related to the complex spinal anatomy and related imaging artifacts, patient factors, tissue contrast and interpretation challenges, as well as cost and accessibility.

The spine has intricate anatomy with various feature sizes leading to significant challenges for MRI [20, 21]. This includes the vertebrae, intervertebral discs, spinal cord, nerve roots, and surround soft tissues. The small dimensions of the cervical spine anatomy, such as the spinal cord in particular, and close proximity to other various structures requires high-resolution imaging to accurately depict the anatomy for clinical purposes such as diagnosis of pathology.

While the ability to image small features can managed by controlling parameters such as matrix size, field of view, and slice thickness to reduce voxel size, this may necessitate longer scan acquisition times. Patient-related factors, such as movement, previous implants, and pathology can then play a large role in image quality. As the voxel size decreases, the importance of the patient's ability to remain still during the scan increases, otherwise, motion artifacts may form. Motion artifacts are inherently formed due to movements such as breathing or swallowing [22, 23], but larger movements, from patients with claustrophobia or severe pain [24] may result in significant motion artifacts. These motion artifacts often include a combination of blurring of sharp edges, ghosting, signal loss, and appearance of undesired strong signals [23].

The ability to achieve optimal tissue contrast when performing spine MRI can be challenging due to the varying properties of the tissues being imaged. This also leads to interpretation challenges for physicians when observing pathology or trauma. Tissue contrast in the images varies significantly and requires careful selection of MRI parameters such as repetition time (TR), echo time (TE), magnetic field strength, and use of contrast agents [25]. When these parameters are optimized, the enhanced contrast provides physicians with an accurate depiction of pathological changes for diagnosis or evaluation [18].

Lastly, cost and accessibility is a major challenge as MRI, with the high cost limiting accessibility, especially in low-resource communities. This is due to the high capital cost of the MRI machine and highly trained personnel that are required to run it. Coupled with the long scan time compared to other imaging modalities such as CT, MRI can be less accessible for spine imaging in some regions.

2.2 Vertebral Detection

For clinical interpretations of spine MRI scans, identifying the specific vertebrae levels can be a critical first step in localizing potential pathology for treatment. Vertebral detection on MRI typically involves a combination of recognizing anatomical features, such as the C2 odontoid process, and counting vertebral levels. The detection and localization of vertebrae levels on MRI can be straightforward given a large high-quality scan; however, it becomes more challenging with smaller and nosier images, often requiring the expertise of radiologists and spinal surgeons for clinical diagnosis and treatment planning.

Recently, algorithmic methods using deep learning techniques have been leveraged for the object detection task, which can be extended to vertebral detection. These advanced methods can automate the process, potentially increasing efficiency and accuracy in identifying vertebrae and aiding in the overall diagnostic process.

2.2.1 Object Detection and Deep Learning

Object detection involves identifying and pinpointing all instances of a specific object within a given FoV [26]. This field has seen significant advancements with the developments in convolutional neural networks (CNNs) and deep learning. These deep learning-based object detectors are now integral to various technologies, ranging from self-driving cars to facial recognition in security systems. Additionally, object detectors are being utilized in medical applications such as detecting pathology in medical imaging and localizing anatomical structures for image-guided surgery [27].

Backbone Architectures in Object Detectors

Backbone architectures are a crucial component within object detectors, enabling the networks to extract features from the input image for detection [26]. These architectures serve as the foundation of the network, processing the raw image data to produce feature maps that highlight important aspects such as edges, textures, and shapes.

Backbone architectures are commonly pre-trained on large-scale datasets, such as ImageNet [28], to expose the network to a vast array of features. This pre-training allows the backbone to learn rich and diverse feature representations, which can significantly enhance the performance of object detectors when fine-tuned on specific tasks.

In addition, backbone architectures have variations themselves, with differing numbers of layers and configurations. These variations allow for performance trade-offs, enabling

users to choose a backbone that balances accuracy and computational efficiency based on their specific requirements. For instance, deeper architectures with more layers can capture more complex features but may require more computational resources and longer training times.

Common backbone architectures include:

- VGG (Visual Geometry Group): Known for its simplicity and depth, VGG architectures (e.g. VGG16, VGG19) use a series of convolutional layers with small receptive fields [29]. Despite their high computational cost, they are widely used due to their straightforward design and effectiveness.
- ResNet (Residual Networks): ResNet architectures (e.g. ResNet50, ResNet101) introduced the concept of residual learning, allowing for the training of very deep networks by mitigating the vanishing gradient problem [30]. This innovation has made ResNet a popular choice for many computer vision tasks.
- EfficientNet: EfficientNet architectures use a compound scaling method to balance network depth, width, and resolution, achieving state-of-the-art performance with fewer parameters and lower computational cost [31]. Variants like EfficientNet-B0 to EfficientNet-B7 offer different levels of complexity and performance.

By leveraging these backbone architectures, object detectors can utilize them as feature extractors to achieve high accuracy and efficiency, making them suitable for a wide range of applications, from autonomous driving to medical imaging.

Two-Stage Object Detectors

Two-stage detectors operate in two distinct phases. The first stage, known as the Region Proposal Network (RPN), generates a set of potential object proposals within an image. These proposals are smaller regions of interest (RoIs) that might contain objects. The second stage then takes these proposals and performs both classification and localization to determine the exact class and bounding box coordinates of the objects.

Due to their two-step process, two-stage detectors are generally more complex and computationally intensive. The separation of the proposal generation and the classification/localization allows for more precise object detection, but it also means that these detectors can be slower and require more computational resources. Additionally, because the classification and localization are performed on the proposed regions rather than the

entire image, these detectors may lack global context, which can sometimes lead to less accurate detections [26].

Despite these challenges, two-stage detectors are known for their high accuracy and are often used in applications where precision is critical. However, their complexity and computational demands make them less suitable for edge devices and real-time applications, where speed and efficiency are paramount. Notable examples of state-of-the-art two-stage object detectors include Faster R-CNN [32] and Mask R-CNN [2].

Single-Stage Object Detectors

Single-stage detectors utilize dense sampling to classify and localize semantic objects in a single shot. This approach involves using predefined anchor boxes or keypoints with various scales and aspect ratios to detect objects. By performing both classification and localization in a single pass through the network, these detectors are less complex and computationally demanding compared to their two-stage counterparts. This efficiency makes them particularly well-suited for edge devices and real-time applications where computational resources are limited [26].

Single-stage detectors achieve this by applying CNNs directly to the entire image, predicting both the class and bounding box coordinates simultaneously. This streamlined process allows for faster inference times, which is crucial for applications such as autonomous driving, surveillance, and mobile devices. Notable examples of state-of-the-art single-stage object detectors include YOLOv8 (You Only Look Once) [33] and RetinaNet [34].

2.3 Multi-Modal Image Registration

Image registration is the process of spatially aligning two or more images. Common reasons for performing image registrationtaken at different times, from different viewpoints, or using different sensors. Specifically, multi-modal image registration is the registration of multiple scans taken with different modalities (eg. MRI and CT). The primary goal of image registration is to transform the images so they share a common coordinate system, enabling meaningful comparison and analysis. In the case of image registration, one image is referred to as the fixed, or reference image, whereas the image being transformed or registered is known as the moving image. Two common forms of image registration include: affine registration and deformable registration.

In medical imaging, particularly for the spine, deformable registration is commonly used. This technique often involves aligning two different scans taken at different times to observe morphological changes. Deformable registration accounts for non-rigid transformations, such as those caused by patient movement or anatomical changes over time, making it particularly useful for tracking disease progression, treatment response, or surgical planning. However, affine registration as a preliminary step is often required to seed initial alignment of the anatomical coordinates of scans prior to deformable registration of the spine.

2.3.1 Image Registration Methodology

To perform image registration between two or more images, four general steps are often conducted [35]:

1. Feature Detection: Control points or features are manually, or automatically detected. These features are salient and distinctive objects such as edges, corners, or closed-boundary regions. In the context of multi-modal spine registration, this could be the detection of features such as the vertebral bodies, intervertebral discs, and/or spinal cord in a MRI or CT scan.

2. Feature Matching: Features detected in the fixed and moving image are matched. These can be matched using various feature descriptors, similarity measures, or spatial relationships. In the case of multi-modal spine registration, detected features such as vertebrae have associated anatomical labels that can be generated and used for feature matching.

3. Transform Model Estimation: The type of transformation model and parameters required are estimated or calculated to register the moving image to the fixed image. This is also known as the mapping function. In multi-modal spine registration, if using an affine registration, the affine transformation matrix is calculated using the detected and matched control points.

4. Image Resampling and Transformation: The moving image is transformed with the mapping function, with interpolation techniques applied where appropriate. This results in the final registered moving image with respect to the fixed image.

2.3.2 Affine Image Registration

Affine registration involves a combination of linear transformations that include translation, rotation, scaling, and shearing to align two datasets. The transformation can be represented by a a matrix equation as seen in Equation 2.1, where $\mathbf{y}$ is the transformed output, $\mathbf{A}$ is the rotation matrix, $\mathbf{x}$ is the input, and $\mathbf{b}$ is the translation vector. The rotation matrix encodes the rotation, scaling, and shear components, whereas the translation vector handles the translation.

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{2.1}$$

Affine registration is particularly useful for aligning images where the relationship between them can be described by these linear transformations. It is applied as a global transformation to the entire input, making it computationally efficient. This efficiency often makes affine registration an initial step before applying more complex registration techniques.

In medical imaging, affine registration is a powerful tool for the initial alignment of scans from different modalities (such as MRI and CT), different time points, or different perspectives. This alignment allows for the comparison and analysis of images, facilitating better diagnosis and treatment planning. Additionally, affine registration is used in other fields such as remote sensing, computer vision, and robotics, where aligning datasets accurately is crucial for further analysis and decision-making.

2.3.3 Deformable Image Registration

Deformable registration is a non-rigid registration that accounts for non-linear transformations such as local distortions and variations between images. This is especially important in the context of medical imaging as anatomical changes, patient movement, or varying imaging conditions between different scans often lead to non-linear transformations. Some common applications of deformable registration in medical imaging include multi-modal image registration, longitudinal studies, and population modelling and statistical atlases [36].

When performing deformable registration, a displacement field or warping (ϕ) is generated to transform a mapping of point locations of moving image to the fixed image in a non-linear function. To generate the optimal transformation, the energy in the objective function (Equation 2.2) must be optimized [36].

$$\mathcal{S}(f, m \circ \phi) + \mathcal{R}(\phi) \tag{2.2}$$

Here, $\mathcal{S}$ quantifies the level of similarity or alignment between the fixed image (f) and the moving image (m). $\mathcal{R}$ is a regularization term to tune the optimization based on any specific properties of the solution that the user may require. Therefore, deformable registration is comprised of three components: a deformation model that generates the transformation mapping, an objective function (as described by Equation 2.2), and an optimization method.

2.4 Translation to Clinical Applications

Multi-modal imaging is often crucial for treatment of complex spinal disorders as different imaging modalities have distinct image contrasts to provide a comprehensive view of the spine. As such, multi-modal image registration is often necessary to align these images accurately. The first step in registration involves detecting and matching features, which is where vertebrae detection becomes essential, yet multi-modal spinal imaging datasets can often have differing FoVs with varying numbers of vertebral levels. Although images can vary in quality and contain artifacts, radiologists and physicians can detect, label, and localize vertebrae of interest based on their training and expertise. However, this process can be time consuming and inefficient.

The goal of this thesis was to develop novel techniques to help automate this vertebral detection process on multi-modal imaging using deep learning techniques. By automatically detecting and labelling the vertebrae in different scans, we can match the detected vertebrae to ensure matching FoVs and use the labelled vertebral levels to seed initial affine registration, which can then improve more complex deformable registration. By creating an automated end-to-end multi-modal registration pipeline, we can ensure accurate and efficient registration to benefit patient diagnosis and treatment. Methods and discussion of the development of tools for vertebral detection and automated multi-modal image registration will be outlined in Chapters 3 and 4.

Chapter 3

Vertebral Detection and Labelling using Deep Learning for Spine MRI Registration

This chapter discusses the development of a deep learning model for vertebrae and spinal cord detection on cervical spine MRI towards the goal of improving multi-modal image registration.

3.1 Introduction

Vertebrae detection and labelling is an essential first step in the automation of spine image registration. The detections can be used to ensure the scans being registered contain the same vertebrae, and for initializing alignment through affine registration. If one image contains the entire cervical spine (C1-C7) and the other image contains a smaller field of view (FoV), or smaller number of cervical vertebrae (e.g. C3-C6), the different image size and spatial features may result in significant registration errors.

We propose the use of an unimodal Mask R-CNN deep learning model for the detection and labelling and segmentation of whole cervical vertebrae and the spinal cord in cervical spine MRI towards improved automated image registration. To train the model, an annotated dataset of vertebrae labels, bounding boxes, and segmentations was generated. Lastly, we leverage the sequential hierarchical structure of the vertebral levels by utilizing a weighted directed acyclic graph (DAG) to filter the detected vertebra and improve

inference outputs from the trained Mask R-CNN model.

3.1.1 Related Works

Prior approaches have automated the detection and labelling of vertebrae in MRI, with classical object detection methods such as histogram of gradients [37], to recent developments utilizing deep learning techniques [38, 39, 40, 41]. Most approaches focus on the detection and labelling of the vertebral body, rather than the whole vertebra (i.e. with the posterior elements). Since the detections are only localizing the vertebral body, the similar size and shape of the vertebral bodies in their respective regions may increase detection and labelling difficulty. In addition, approaches are often developed for use in a fixed FoV of a specific region of the spine, such as the cervical or lumbar spine only. This is primarily due to the high costs associated with MRI, leading to images being targeted to specific regions, rather than the whole spine. While these approaches do have some variations in the FoV, such as training on images with T10-L4 and T12-S1 FoVs [40], they are not intended for use with even smaller FoVs that only encompass the lumbar spine (eg. L1-L4).

Alongside the literature on vertebrae detection in MRI, significant advancements have also been made to the detection of vertebrae in CT scans [42, 43, 44, 45]. Detecting vertebrae in CT scans is less challenging than in MRI because of more consistent image intensities (due to the source of contrast), bone in CT having excellent contrast against soft tissue, CT images typically having higher image resolution, and scans containing larger FoVs because of the imaging speed and cost of the modality. Approaches to vertebrae detection in CT have ranged from methods using statistical shape models [46, 47], deformable atlases [48], to deep learning in recent years. Suzani *et al.* [49] was one of the first using a feed-forward neural network to detect and localize vertebrae centroids using intensity based features extracted from a CT image.

Literature on deep learning models utilized for vertebrae detection, in both MRI and CT, primarily use a multi-stage object detector architecture approach. Multi-stage object detection architectures tend to utilize multiple models or modules to break down the detection task into sequential steps for greater performance. Often, the first stage extracts features or a region of interest (RoI) from the image, and the second stage classifies and localizes the objects. In contrast, single-stage object detection architectures directly predict bounding boxes and class probabilities as a regression problem. This leads to multi-stage detection models being more accurate, with the trade-off of requiring more computational power.

Prior works using multi-stage object detector architectures include Zhao *et al.* [50], who

proposed a hierarchical self-calibration detection framework to recognize vertebrae in arbitrary MRI FoVs. This model utilized a hierarchical detection network to propose regions containing vertebrae and a self-calibration recognition network to classify the proposed regions to bounding boxes. Message passing was used to re-calibrate class probabilities to ensure hierarchically enforced labelling. Windsor *et al.* [38, 51] proposed a novel method of detecting vertebrae in 3D within the whole spine by first utilizing a key point detection module for localization of vertebral body corners and the centroid, rather than a region proposal network. A secondary labelling module leveraging an appearance and context network was then used to leverage the appearance and spatial configuration of the detections to ensure sequential hierarchy of the vertebrae. Later, Zeng *et al.* [39] proposed FORCE, a method to to detect vertebrae within a 2D slice to tackle vertebral appearance and FoV challenges. FORCE utilized three modules; a region proposal network module for RoI extraction, an Eigenmap-based feature similarity regularization (FSR) module to promote similar vertebrae to be closer in latent space representation for higher feature discrimination, and a cumulative sparse representation module to preserve historical features from being erased for bounding box and class predictions. Zhao *et al.* [40] trained a feature sequence-based detection framework, consisting of three modules for 2D vertebrae detection. This includes a local feature extraction module, a discriminative sequential image description module to generate representative sequences to describe the vertebrae, and a spinal pattern exploitation module to capitalize on the hierarchical pattern of the vertebrae sequence.

While these complex models are successful in detecting, localizing, and classifying vertebrae, they are developed using fixed, non-varying FoVs of entire spinal regions. The fixed FoVs are incompatible with the central clinical need being addressed in this investigation, specifically image guided therapy in the spine, that will typically have images with varying and small FoVs focused on the levels to be treated. To detect vertebrae in small, varying spine FoVs, the Mask R-CNN architecture [2] was considered a candidate for MRI vertebrae detection and labelling. This choice was due to its cutting edge object detection capabilities and ease of implementation. The Mask R-CNN utilizes a two stage architecture design (Figure 3.3), where stage one is a backbone with a region proposal network to extract potential regions of objects, and stage two utilizes three heads to classify, predict refined bounding boxes, and output instance segmentation masks. Recently, Wang *et al.* [52] found success in characterizing the performance of a Mask R-CNN model for detecting vertebrae from maximum intensity projection images from CT. Additionally, studies have used the Faster R-CNN [32] architecture as a comparative baseline for vertebrae detection and labelling [50, 39, 40]. The Mask R-CNN architecture is an improvement to the Faster R-CNN architecture with an added instance segmentation head, leading to

the investigation into the performance of the Mask R-CNN for cervical vertebrae detection and labelling.

Another feature found in literature describing vertebrae detection and labelling methods includes the leveraging of the sequential hierarchy of the vertebrae. Zhan *et al.* [53] proposed the use of a distinct vertebra detector (eg. C2 vertebra) and bundle vertebrae detector (eg. C3-C7 vertebrae) to leverage the hierarchical structure of the spine. This utilized the detected anchor vertebra to subsequently label the bundle vertebrae detected. This strategy emulates the methods of a experts as they generally use an anchor vertebra to orient themselves to the vertebral position, prior to labelling the remaining vertebrae. Another method to leverage the hierarchical structure of the vertebrae is to use a graph data structure, or a graph convolutional network (GCN). A graph is a collection of nodes (vertices) and edges connecting pairs of nodes. It can represent various systems such as social networks and the links between people. In addition, edges of a graph can be directed and weighted. In contrast, a GCN is a type of neural network designed to adopt the graph-structured data. This is done by extending the concept of convolution from traditional grid data, like images, to graphs for tasks like node classification and link prediction. Forsberg *et al.* [41] used two separate CNN vertebrae detectors, one for general vertebrae detection and one for C2 or S1 detection, as inputs to a parts-based graph model. The graph combined the detections and labels and outputs a configuration that follows the sequential nature of the vertebrae labels and spatial location of each detection. Later, Chang *et al.* [54] used a spatial GCN to ensure global spatial relations between the vertebrae are established, followed by a label attention network to generate labels and handle mis-classification on 2D slices. Recently, Klein *et al.* [45] proposed VertDetect, a multi-stage approach to detect, localize, label, and segment the whole spine in CT volumes. It utilizes three branches including: 1) a detection branch, 2) classification branch, and 3) a segmentation branch. The vertebrae detection branch utilized a modified 3D ResNet-50 backbone and feature pyramid network for feature map extraction followed by multiple convolution layers to output vertebral body bounding boxes, vertebral body centroid offsets, and vertebral body centroid heatmap. The classification branch utilized a GCN to leverage the hierarchical nature of the vertebrae to label the detected vertebra. The segmentation branch utilized the centroid heatmap to seed segmentation of vertebrae bounding boxes previously localized. Lastly, a post process label adjustment was performed using a graph to ensure all labels were sequential. The success of graphs and GCNs to enforcing sequential hierarchy of vertebrae in literature demonstrates that investigation into the application and performance of graphs is required when developing a vertebrae detector. In this investigation, we only use a graph data structure to improve labelling and enforce sequential hierarchy instead of a GCN, which would increase model and pipeline complexity without increased

gains in performance [45].

3.2 Methods

To assess the performance of the Mask R-CNN model for cervical vertebra and spinal cord detection, a novel labelled dataset of cervical spine MRI with different FoVs was generated based on the open-source Spine Generic dataset. The model was then trained on the generated dataset, with predicted results filtered with a graph. Integration into a multi-modal spine image registration pipeline and registration experiments were performed in Chapter 4.

3.2.1 Dataset Generation

This study utilized the open Spine Generic dataset to generate a novel annotated dataset. It consists of anatomical T1-weighted (T1w) and T2-weighted (T2w) images, and diffusion weighted (DW) images from 267 healthy subjects imaged from 42+ centers worldwide [1]. The images were taken using the Spine Generic protocol [55] to ensure quantitatively consistent imaging is captured for spinal cord research. A sagittal acquisition was employed to efficiently cover the superior-inferior (SI) axis, minimizing the number of slices needed to encompass the spinal cord [55]. In addition, the T1w and T2w scans were taken with an isotropic resolution with 3T MRI machines from Phillips, Siemens, or GE.

The original T1w images of the Spine Generic dataset (Figure 3.1 a)) contained the entire skull and cervical spine, ending at various vertebra in the thoracic region. In comparison, the T2w images (Figure 3.1 b)) did not contain the entire skull, but still included various vertebra into the thoracic region. The spine was also aligned in the SI direction of both scans. The T1w images have a voxel size of 1 mm^3 with images being approximately 192x260x320 voxels (RAS). The T2w images have a voxel size of 0.8 mm^3 with images being approximately 64x320x320 voxels (RAS).

For the purposes of this study, only the 2D mid-sagittal slice of the T1w and T2w spine anatomy were utilized for training the model. The use of the anatomical mid-sagittal slice was chosen due to the unique anatomy of the vertebrae visualized in that 2D slice. Specifically, C2 has the unique odontoid process extending the length of the vertebral body with C1 as a ring surrounding the C2 odontoid process as seen in Figure 2.3. In addition, the length of the spinous processes of the C2-C7 vertebrae is typically largest at the mid-sagittal slice, with C7 having the longest spinous process of the cervical spine as seen in



Figure 3.1: A single subject T1w and T2w MRI from the spine generic dataset [1]. a) T1w image containing entire skull, cervical spine, extending into the thoracic spine. The image has an isotropic voxel size of 1 mm^3 and has a voxel resolution of approximately 192x260x320 (RAS). b) T2w image containing the cervical spine, extending into the thoracic spine. The image has an isotropic voxel size of 0.8 mm^3 with a voxel resolution of approximately 64x320x320 (RAS)

Figure 2.1. Lastly, a mid-sagittal slice was used due to the Mask R-CNN model being developed for 2D object detection [2]. The mid-sagittal slice of volume from the Spine Generic dataset was found to be approximately the anatomical mid-sagittal slice of the spine and spinal cord, therefore sufficient for model training.

The raw Spine Generic dataset was then semi-automatically processed to generate ground truth data for the supervised training of the Mask R-CNN model. This included generating vertebrae and spinal cord segmentations, bounding boxes, and class labels within the MRI volumes. Then, the mid-sagittal slice of the generated ground truth data was extracted and used for model training. This was done algorithmically, utilizing the Spinal Cord Toolbox and 3D Slicer platforms, with some manual cleaning. Figure 3.2 outlines the dataset generation workflow.

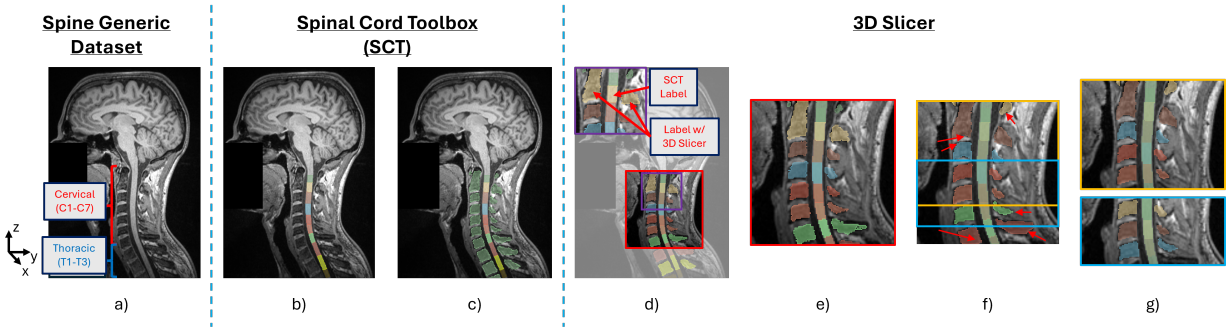


Figure 3.2: Dataset generation workflow used to obtain ground truth annotations of the T1w Spine Generic Dataset. A similar process is used to generate the T2w dataset. a) The raw Spine Generic T1w MRI images of healthy subjects. b) Initial segmentation and labelling of the spinal cord using SCT Deepseg with different colours representing different instance segmentations of the vertebral levels on the spinal cord. A semantic segmentation of the whole spinal cord is also generated. c) Registration of PAM50 template vertebral column using SCT. d) Relabelling registered PAM50 template into individual vertebra and incorporation of spinal cord segmentation in 3D Slicer. e) Cropping to the cervical spine (C1-C7) after quality checks. f) Manual segmentation clean-up and removal of any thoracic vertebra segmentations and labels visible highlighted by red arrows. g) Generate various FoV windows of the cervical spine with a minimum of 3 vertebra.

Spinal Cord Toolbox

Spinal Cord Toolbox (SCT) was utilized to generate ground truth annotations from the Spine Generic dataset automatically. SCT is an open source platform developed for the analysis and processing of spinal cord MRI with standardized templates and analysis procedures [56]. While SCT is tailored for spinal cord analysis, it also provides robust tools for spine analysis, including the automated generation of a vertebral column segmentation and labels of vertebral levels.

To obtain the cervical spine segmentation and vertebral level labels, the SCT Deepseg algorithm was first utilized. This algorithm utilizes a modified 2D U-Net convolutional neural network to generate a semantic segmentation of the spinal cord [57]. The raw Spine Generic scan (Figure 3.2 a)) was processed to obtain a spinal cord segmentation. This was then split into smaller instances along the spinal cord that were labelled based on the adjacent vertebral levels as seen in Figure 3.2 b). Next, the spinal cord segmentation and instance segmentations of the vertebral levels on the spinal cord were used to register the PAM50 template, an unbiased multi-modal MRI template of the whole vertebral column

[58], to create the ground truth semantic segmentation of the spine as seen in Figure 3.2 c).

3D Slicer

3D Slicer is an open source platform for medical image computing [59]. In this study, 3D Slicer was used to generate individual vertebra instance segmentations from the PAM50 spine semantic segmentation. This was done by splitting the semantic segmentation into individual islands (vertebra) in 3D Slicer by leveraging the disjointed nature of the vertebrae within the semantic segmentation. The vertebral level segmentations on the spinal cord from SCT's Deepseg algorithm were then used to label the individual vertebra segmentations based on the nearest label centroid to each vertebra island centroid. The semantic spinal cord segmentation was also combined with the instance vertebra segmentations at this point, resulting in a final segmentation with the individually labelled vertebra with their corresponding instance segmentations, and spinal cord semantic segmentation as seen in Figure 3.2 d). Quality checks of the images were then performed to ensure segmentations and labels were generated correctly for training ground truth. Of the 267 T1w and T2w images from Spine Generic, only 149 T1w and 100 T2w images were used due to substantial segmentation and/or labelling errors found.

The MRI volumes and their accompanying segmentations were then cropped to only include the cervical spine. This was done by utilizing the generated segmentations and cropping the superior-inferior (SI) direction of the image to the top-most and bottom-most extent of the segmentations with an additional spacing of 5 voxels. The additional spacing was used to compensate for any minor segmentation errors resulting in the vertebrae being slightly cut off during the cropping process. The anterior-posterior (AP) and medial-lateral (ML) directions of the image was also cropped based on the extreme boundaries of the segmentations with additional spacing as well. The additional spacing in the AP and ML directions was a function of the image size subtracted by the voxel width of the between the extreme boundaries of the specific axis, divided by 3. This ensured that the images would contain the entire neck and some of the air gap as seen in Figure 3.2 e), which was found to be similar to the FoV in the clinical data used to verify the model outlined in Chapter 4.

A 2D mid-sagittal slice of the cropped cervical spine volumes and processed segmentations were then extracted for the training of the model. Additional minor manual segmentation edits were performed to clean up the 2D mid-sagittal segmentations of the vertebrae to ensure robust ground truth segmentation masks, which were used for bounding box coordinate generation when training (Figure 3.2 f)).

The 2D mid-sagittal slices and the segmentations were then cropped into various FoVs in the SI direction with a minimum of three vertebra within a window, e.g. C2-C5 and C4-C6 (Figure 3.2 g)), resulting in a total of 1341 and 900 images for T1w and T2w, respectively (9 images from each subject). Note that C1 and C2 were always counted as one vertebra (C1/C2) within the context of the three vertebrae window. This was due to the anatomical position of C1 relative to C2 leading to C1 being visible if C2 is present. The generation of various FoVs was conducted to enlarge the dataset and ensure that the model does not implicitly expect all cervical vertebrae to be present in all inputs. This also provides the model with information on smaller FoVs and sizes of vertebrae in the image, which can be clinically relevant.

Lastly, a similar cropping into various FoV was done with the cropped cervical spine volumes with their associated segmentations. These volumetric segmentations were not manually cleaned and were generated for use for FoV cropping and registration experiments outlined in Chapter 4.

We train the Mask R-CNN model using the generated annotated dataset of small and varying FoVs of the cervical spine in 2D slices to perform instance vertebrae segmentations and detection of vertebrae bounding boxes. In addition, we semantically segment the spinal cord and localize the spinal cord bounding box.

3.2.2 Mask R-CNN Vertebra Detector Model

The Mask R-CNN [2] is a cutting edge object detection architecture and was evaluated in this study for its performance in detection of cervical vertebra. The architecture is capable of instance segmentation and adopts from the Faster R-CNN architecture [2] as seen in Figure 3.3. It consists of a region proposal network (RPN) that often utilizes a ResNet backbone feature pyramid network (FPN) to generate a feature map. The RPN proposes regions of interest (RoIs) which are extracted from the feature map with a RoI align. The RoI align process extracts features from the proposed RoIs using bilinear interpolation to avoid misalignments caused by quantization using the RoI pooling method of the Faster R-CNN [2]. The extracted feature maps are then used as inputs into a segmentation, bounding box generator, and classification branch. The bounding box and classification branches utilize a shared fully connected (FC) layer followed by a split into separate FC layers. The classification branch uses a final softmax layer to classify the RoIs. In contrast, the bounding box branch has a final FC layer to regress the bounding box coordinates and refine the initial RoI proposals coordinates. Lastly, the segmentation (mask) branch uses convolutional layers to predict the segmentations of the object within the RoI.

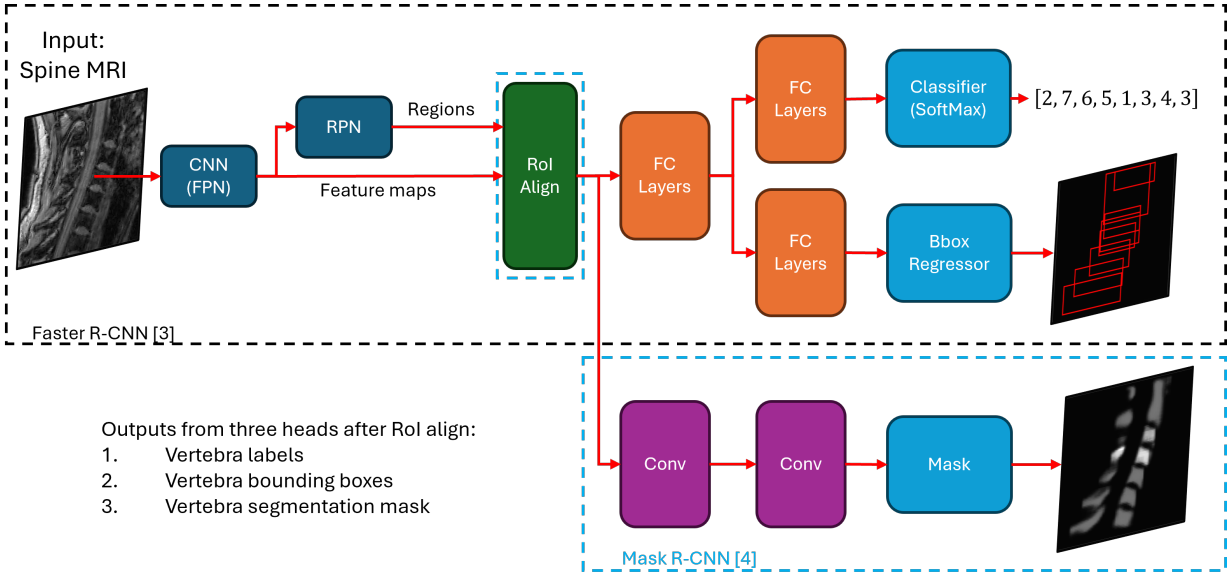


Figure 3.3: Mask R-CNN architecture for object detection and instance segmentation. Feature maps are extracted from the input image using a ResNet backbone FPN, with object RoIs proposed by an RPN. A RoI Align process then extracts the regions from the feature maps, followed by the classification, bounding box regression, and instance segmentation (mask) branches for their respective tasks.

The predicted bounding boxes and class labels can be utilized for the detection of vertebra for image FoV cropping. The predicted segmentations can be incorporated into clinical workflows, such as generation of 3D models for improved navigational planning when registered with an accompanying image and segmentation pair.

For this study, the Pytorch Torchvision implementation of the Mask R-CNN architecture was utilized (Torchvision v0.18.0). Two models were trained, one for each MRI image sequence (T1w and T2w). The models utilized a ResNet50 FPN backbone pre-trained on the COCO 2017 dataset (using z-score normalization) for feature extraction within the scans. In addition, a multi-task loss, L, is used for model training on each sampled ROI by summing the following: $L_{classifer}$, L_{box}, and L_{mask}. $L_{classifer}$ is a categorical cross-entropy loss responsible for penalizing the model during the classification task of labelling the bounding boxes proposed in the ROI [32]. L_{box} is a regression loss of the bounding box coordinates utilizing a smooth L1 loss [32]. Lastly, L_{mask} is an average binary cross-entropy loss comparing the binary mask generated by the model to the ground truth.

Data and Implementation

The dataset was randomly split for training, validation, and testing 70%:20%:10%. Data was stratified by subject to ensure no data leakage due to the generation of multiple vertebrae FoVs from individual subjects. This equated to 106:29:14 subjects (954:261:126 images) and 70:20:10 subjects (630:180:90 images) for T1w and T2w, respectively. Since the 2D mid-sagittal slice was used, the single slice was treated as a 2D image in the training process. Further data augmentations were applied online to expand the training set and improve generalization as follows:

- Min-max normalization: Images were normalized to values between 0 and 1 using min-max normalization as outlined in Equation 3.1. This was done to ensure MRI training inputs had consistent voxel values, as differing MRI scan protocols can result in differing voxel value ranges.

$$\mathbf{I} = \frac{I - \min(I)}{\max(I) - \min(I)} \tag{3.1}$$

- Resampling: Images were resampled to a voxel size of 1mm^3 to ensure consistency when training the models. This results in resampling only the T2w data as it has a voxel size of 0.8mm^3

- Random affine transform: A random affine transform was utilized to apply shear to the image, resulting in changes to the spine curvature. Since the Spine Generic dataset only consisted of healthy subjects, all cervical spines in the dataset have a lordotic curvature. The random affine transform utilizes no rotation, random translation of 10% of the image size in the i and j directions, random scaling between 75% and 125%, and random shear in the i and j direction ± 20 voxels using bi-linear interpolation.

- Scaling: Scaling was performed to ensure the input image would encompass the entire image grid of 225x225 for subsequent transformations. For images with anisotropic image ratios, the larger image axis would be scaled to 225 pixels.

- Random Rotation: Random rotations were applied to augment the orientation of the vertebrae. The image was randomly rotated ± 10 degrees. In addition, the image grid was expanded to ensure the entire randomly rotated image was preserved.

- Elastic Transform. An elastic transform was applied to the images to transform the morphology of the image with a see-through-water-like effect. This was applied to

degrade the image and similar to potential artifacts in clinical data. The elastic transform was applied using bi-linear interpolation with an alpha of 50 and sigma of 5.

- Random Adjust Sharpness: A random sharpness adjustment was made to degrade the edge contrast within the images. This was done to mimic lower quality clinical scans. Sharpness was reduced to a sharpness factor of 0.5 with a 50% probability.
- Gaussian Blur: A random Gaussian blur was applied to further degrade the input images uniformly to mimic lower quality clinical scans. A kernel size of 3 with a random σ.
- Padding: Padding was applied to ensure all input images were of the same size for training. Zero padding was utilized, resulting in an 250x250 image.

Validation and testing augmentations only included normalization, resampling, scaling, and padding to ensure images and features were similar to the training data in size. A seed value of 10 was selected to standardize the data split and augmentations between trained models.

The model was trained using an ADAMs optimizer with a learning rate of $1e^{-4}$, a batch size of 5, and utilized a confidence threshold of 0.5. The models were trained on a single NVIDIA 4070S GPU with 12GB of VRAM till convergence of the multi-task validation loss. The repository was programmed using Python v3.10.14, using PyTorch v2.3.0 with CUDA v11.8 and Torchvision v0.18.0. The T1w model was trained for 100 epochs and the T2w model was trained for 50 epochs. In addition, the T2w model utilized transfer learning and was initialized using the pre-trained weights from the T1w model. Training and validation loss curves can be found in Appendix A.

Post-Inference Filtering using Graph Structure

After inference, the predictions were filtered using a confidence score threshold of 0.6 to ensure a balance of correct and incorrect predictions. This includes undetected vertebrae, or false negatives (FNs) and extra detections, or false positives (FPs). Investigation of the inference score threshold choice can be found in Section 3.3.3.

After predictions were filtered with a confidence score threshold, a filtering technique was required to ensure each vertebrae only has one associated detection from the large amount of detections from the model. This was done by leveraging the sequential hierarchical order of the vertebrae by adopting a similar method to [45], where the confidence

score of the k filtered predictions were then used to generate of a weighted directed acyclic graph (DAG). By utilizing the prediction labels and scores, we can leverage the hierarchical structure of the vertebral levels to generate the weighted DAG. Each cervical vertebra is numbered 1-7, represented by N, and is based on the position in the superior-inferior (SI) direction from top to bottom. Each detection, consisting of a bounding box, mask, and label, corresponded to a node (V) and the model confidence score of that prediction was used as the edge weight (E).

To correct FP detections where the label is incorrect, but with correct detections, additional nodes are generated as lower confidence predictions. If the confidence score of a prediction was less than 90%, two neighbouring nodes were generated, consisting of the same bounding box and mask, but with the labels representing the neighbouring vertebra superior and inferior the original prediction. For example, if the model predicted C4 with a confidence of 80%, two additional predictions, or nodes, of C3 and C5 would be appended with the same bounding box and mask as the original C4 prediction. The two neighbouring nodes were also given a 50% scaled confidence score relative to the original prediction for use as the edge weight, resulting in an edge weight of 40%. A starting top (T) and ending bottom (B) node was added to the graph that connect to the superior-most and inferior-most nodes.

An adjacency list was created to represent the weighted DAG, $G(V, E)$. This was constructed using the model predictions and generated neighbour nodes, as seen in Figure 3.4, using the following handcrafted rules:

1. The bounding box centroid location (SI axis) of the current node must be less (more inferior) than the centroid location of the previous node to enforce vertebrae spatial positional hierarchy.

2. The bounding box centroid location must be greater than 8 voxels and less than 20 voxels from the centroid location of the previous node to prevent a double detection of the same vertebra.

    (a) In the case of the C1 and C2 in the graph, the bounding box centroid location must be greater than 4 voxels and less than 20 voxels. This is less constraining compared to other vertebrae due to the anatomical location of C1 relative to C2, resulting in overlapping bounding boxes and closer centroids.

3. The current node label has to sequentially next in the vertebral level hierarchy (ie. C2 $\rightarrow$ C3).

Each node (V) is represented by the predicted label $n_k^{(N)}$. The edge weight is the prediction confidence from the model represented as $w_k^{(N)}$. Here, k represents the k^{th} prediction from the model and N represents the vertebral level label from the model prediction or neighbour node generation. The edge weight of the top and bottom nodes are zero ($w_T = w_B = 0$).

Once the weighted DAG adjacency list is constructed, the graph filtering and relabelling of the predictions is performed by solving the longest path of the graph from top (T) to bottom (B) and vice versa. The path with the highest sum of edge weights were used as the final filtered detections. The construction of the longest path from top to bottom and bottom to top was performed to ensure the weighted DAG does not prematurely stop traversing and filter out too many predictions when starting from a certain position. In addition, since the T and B nodes connect to the superior and inferior-most vertebrae, respectively, this also ensure that the longest path takes into account starting with a generated neighbour node.

Figure 3.5 demonstrates a simplified example of the graph filtering and generation of neighbour nodes if the model predicted C3 (99%) on the C3 vertebra, C4 (60%) on the C3 vertebra as a double detection, C4(99%) on a C4 vertebra, and C4 (80%) on a C5 vertebra as a mislabelled detection. The anatomical location of the ground truth C3, C4, and C5 vertebrae are outlined using the dashed boxes. The orange neighbour nodes are generated when the confidence score from the model is below 90%. These neighbour nodes copy the spatial information from the original prediction (mask and bounding box), but are relabelled to the adjacent vertebrae superior and inferior to the original prediction. In addition, the edge weight is reduced by 50%. The weighted DAG adjacency list is then calculated using the stated rules, with valid paths: 1) $T \rightarrow n_0^{(3)} \rightarrow n_2^{(4)} \rightarrow n_3^{(5)} \rightarrow B$, and 2) $T \rightarrow n_1^{(3)} \rightarrow n_2^{(4)} \rightarrow n_3^{(5)} \rightarrow B$. The longest path when summing the edge weights is path 1. This results in the weighted DAG filtering out prediction 1 and the generated neighbour nodes, instead using the generated neighbour node of prediction 3. A similar process is done from B to T, resulting in the same longest path.

3.2.3 Evaluation Criteria

Experiments were performed to validate the performance of the Mask R-CNN model in cervical vertebra detection.

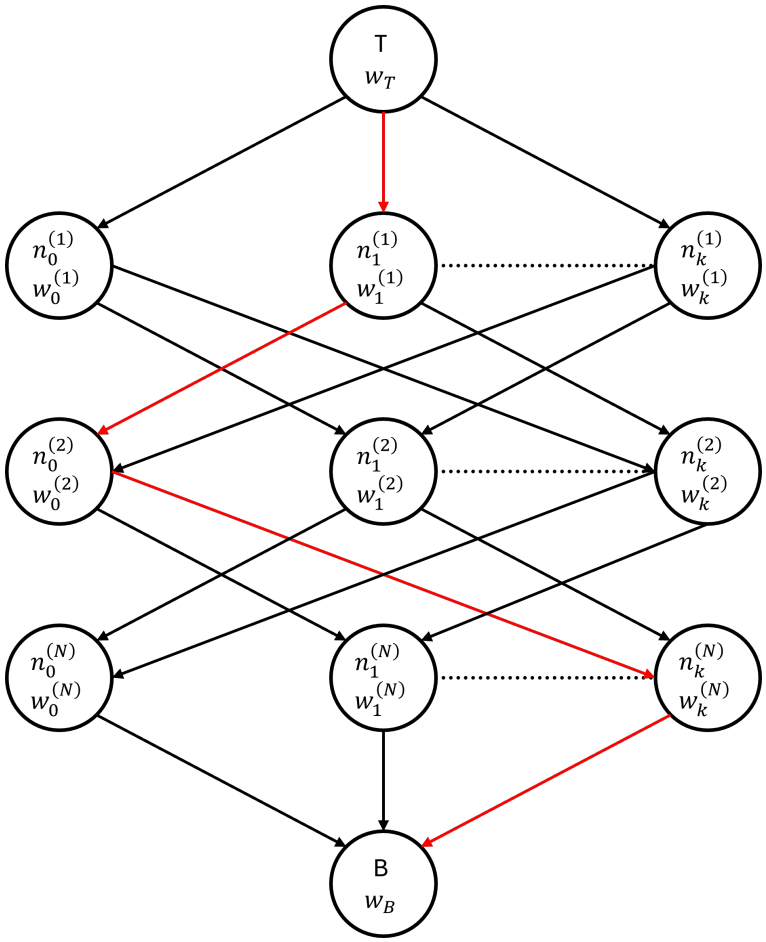


Figure 3.4: General representation of the post-inference graph structure. Each node consists of the predicted label of a vertebra detection, n, and edge weight, w, representing the model confidence of the detection. N represents the vertebrae label and k represents the k^{th} prediction. Each row represents a vertebral level, or additional generated neighbour nodes with different vertebral level labels, with each column representing a prediction from the model. The the nodes from the same prediction, k, cannot connect as they are either the model prediction, or neighbour node with the same mask and bounding box values. The T and B nodes correspond to the top and bottom of the detections and have weights of zero. The nodes of the weighted DAG are connected based on three main rules. The red path is an example of the longest path from T to B when solving the weighted DAG.

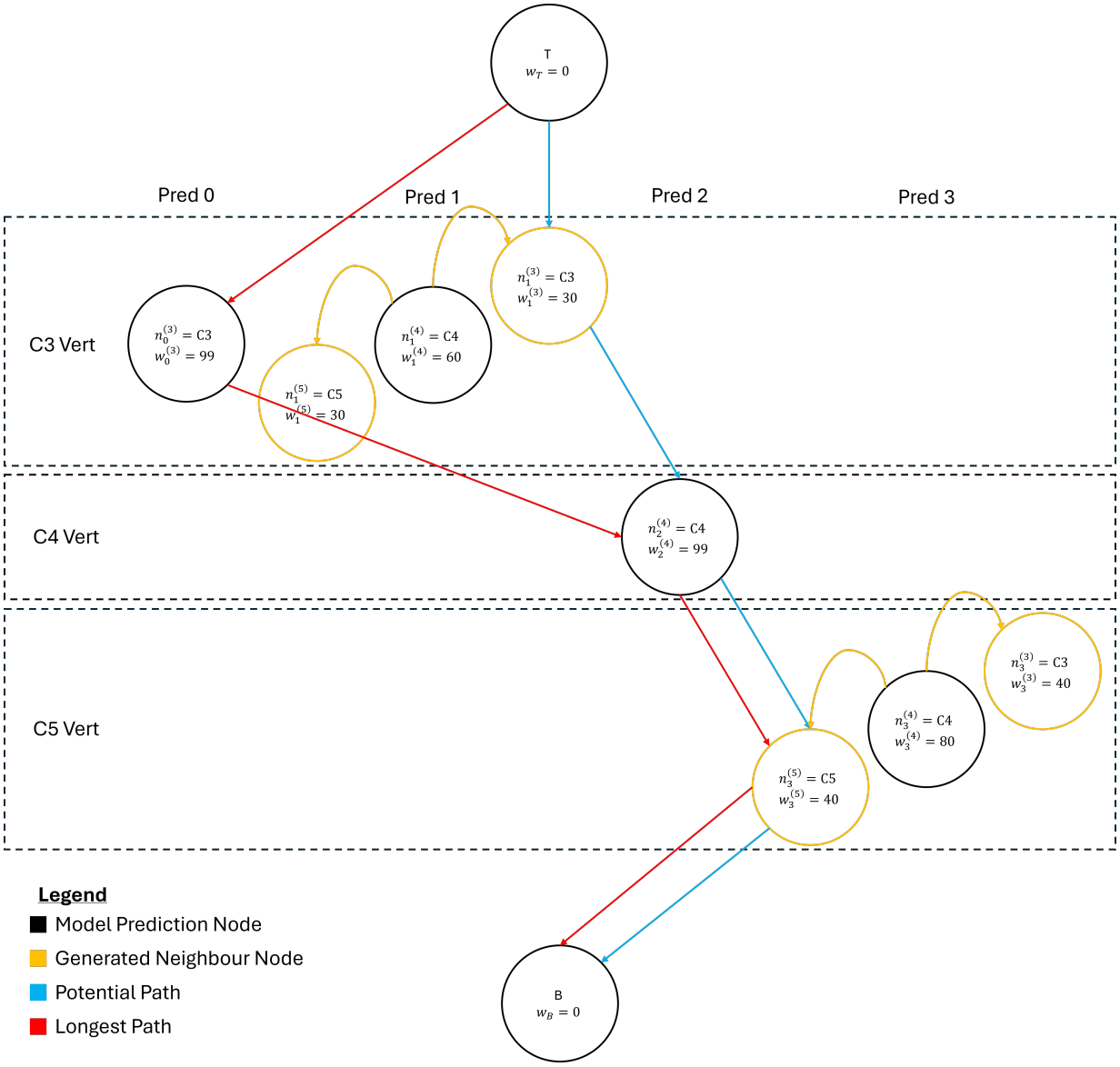


Figure 3.5: Example of the graph filtering four inference predictions with ground truth vertebrae and anatomical location of each prediction outlined by the dashed box from top to bottom. Orange nodes are generated neighbouring nodes due to the prediction having a confidence score less than 90%. The longest path is then calculated to filter and relabel predictions, highlighted in red.

Qualitative Performance

Qualitative evaluation of the vertebra detection model was conducted to demonstrate the accuracy and robustness of the Mask R-CNN model. Specifically, challenging images with different FoVs with and without landmark vertebra were selected from the test set. Predicted bounding boxes, masks, and labels of the vertebrae in T1w and T2w images display the performance of the model.

Quantitative Performance

Quantitative evaluation of the vertebra detection model was conducted using six metrics:

1. Mean average precision with a 75% intersection over union (IoU) threshold (mAP_{75}): A widely used challenge metric to evaluate state-of-the-art object detection models, including Mask R-CNN [2], by highlighting the model's robustness in detecting and labelling with a high IoU threshold. It calculates the average of the Average Precision (AP) values across all classes predicted by the model. AP itself is derived from the area under the precision-recall curve for a single class and IoU threshold, in this case 75%. Precision measures the accuracy of positive predictions, derived from Equation 3.2. In contrast, recall assesses the ability to identify all relevant instances, calculated using Equation 3.3. The IoU threshold determines if a predicted bounding box detection is a true positive, based on the percent overlap of the prediction to the ground truth.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3.2}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3.3}$$

2. Localization error (LE) [mm]: The mean absolute error (MAE) between the predicted and ground truth bounding box centroids [42] based on the voxel size. This measures the localization performance of the bounding boxes generated during vertebrae detection. We resample the voxel size to 1 mm^3 for this study, resulting in a LE unit of millimetres [mm]. Note that this metric only evaluates the distance of predicted vertebrae to the ground truth, and does not consider undetected vertebrae.

3. Identification rate (IDR): A measure of the detection performance of individual vertebra by the model. It is a percentage of vertebra correctly detected and labelled compared to the total vertebra in the image. A vertebra is considered correctly detected and labelled if nearest ground truth label is the same and the predicted bounding box LE is less than 20 mm [42].

4. Image recognition accuracy (IRA): A percentage of images with all vertebra correctly detected and identified without any vertebrae undetected or having extra vertebrae detections. A challenge metric studies have used to evaluate spine vertebra detection models [50, 39].

5. Limiting FoV accuracy (LFA): A proposed metric examining the percentage of images with the two limiting vertebra of the FoV detected, i.e. top-most and bottom-most vertebra detected. This is a proposed metric to quantify the clinical relevance of the detections as the task of FoV cropping towards image registration can be performed with just the detection of the two limiting vertebra in the image.

6. Dice similarity coefficient: A common segmentation metric used in medical calculated as the magnitude of the volume of overlap of predicted and ground truth segmentations over the average of the prediction and ground truth volume magnitudes.

7. F1 Score: A common classification metric that calculates the harmonic mean of precision and recall.

In addition, the AP of each vertebra (IoU thresholds from 0.5 to 0.95 with 0.05 increments) and confusion matrices were observed as a supplementary value to determine model performance on the spinal cord and differing vertebral levels in the cervical spine. This also allowed for insight on the performance of the model when many vertebra are very similar in shape and size. Ablation and comparative experiments will also be discussed.

3.3 Results and Discussion

The models were successfully trained to detect cervical vertebra in T1w and T2w MRI images, and evaluated on the held test set unseen during training and validation.

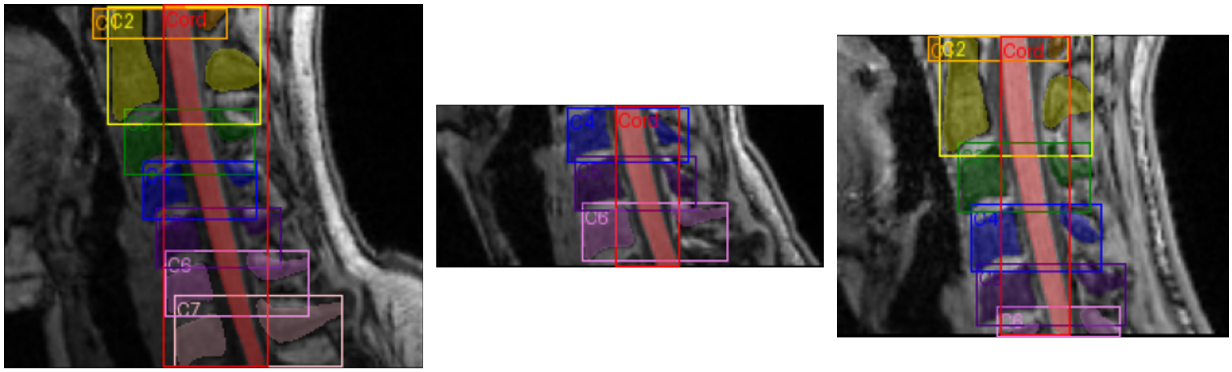


Figure 3.6: Example inference outputs of the T1w Mask R-CNN model for cervical vertebra and spinal cord detection show excellent vertebra detection performance with varying vertebra FoV. Each inference output is from a different subject. a) Detection of C1-C7. b) Detection of C4-C6. c) Detection of C1-C6 with C7 extra detection.

3.3.1 Qualitative Evaluation of Mask R-CNN Model

Figure 3.6 and Figure 3.7 show examples of the inference results processed with the graph filter of the T1w and T2w images, respectively. We demonstrate that the Mask R-CNN model was able to successfully detect and label cervical vertebrae and the spinal cord in the T1w and T2w images with both models showing the ability to detect and label all cervical vertebrae and the spinal cord as seen in Figure 3.6 a) and Figure 3.7 a). In both cases, the model was able to generalize and detect vertebrae in varying FoVs, even in the challenging case such as C4-C6 (Figure 3.6 b)) in the T1w model and C3-C6 (Figure 3.6 b)) in the T2w model. This is difficult due to the lack of distinct landmark vertebrae, such as C2 in the image, and the nearly identical vertebrae size and shape. Lastly, Figure 3.6 c) and Figure 3.7 c) show extra detections of partially cut vertebrae from the data generation process due to the curvature of the spine. However, these extra vertebrae detections are filtered correctly using the graph and have anatomically correct labels.

3.3.2 Quantitative Evaluation of Mask R-CNN Model

Table 3.1 outlines the evaluated metrics of the T1w and T2w model's ability to correctly identify vertebra and generate robust bounding boxes and segmentations for downstream registration improvements. The models performed exceptionally well as observed by the

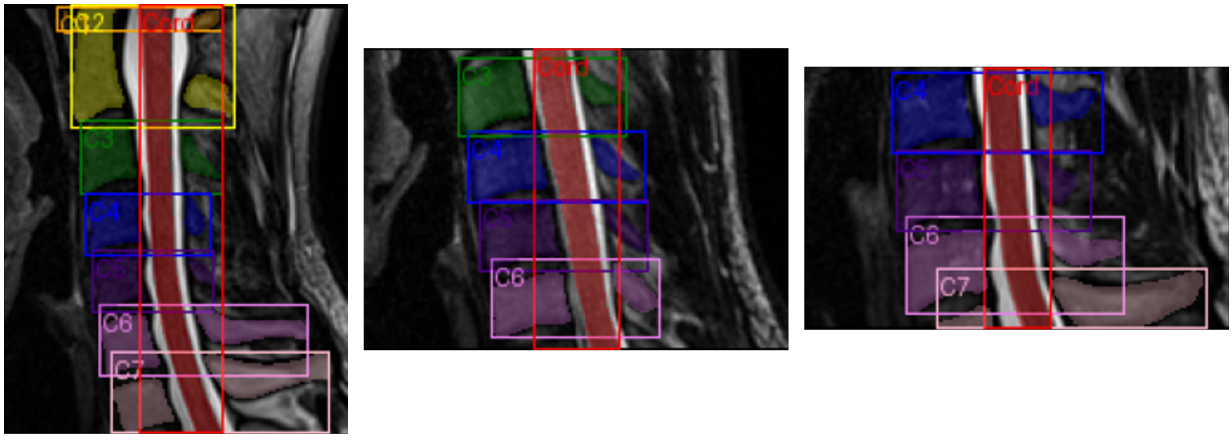


Figure 3.7: Example inference outputs of the T2w Mask R-CNN model for cervical vertebra and spinal cord detection show excellent vertebra detection performance with varying vertebra FoV. Each inference output is from a different subject. a) Detection of C1-C7. b) Detection of C3-C6. c) Detection of C4-C6 with C7 extra detection.

mAP_{75} value of 0.919 and 0.867 for T1w and T2w images, respectively. The low LE and high Dice coefficient demonstrates the models ability to localize bounding boxes with minimal error and generate robust segmentations of the vertebrae. The generation of good bounding boxes for its centroid is vital for cropping mismatched FoVs and seed initial affine registration towards improving deformable registration. The segmentations are used to quantify the registration performance and can be used for downstream clinical tasks. IDR and IRA metrics resulted in 96.0% and 73.1% for the T1w images, and 95.0% and 84.4% for the T2w images. The high IDR for both models indicated that the model, with the use of post-process graph filtering, was able to detect almost all vertebrae present in the images, resulting in a low number of undetected vertebrae. In contrast, the lower IRA observed in both models indicates that there are a high number of extra FP vertebrae detections as seen in Figure 3.6 c) and Figure 3.7 c). The F1 score shows that the models were successful in classification of the vertebrae, with high precision and recall. Lastly, the LFA metric demonstrates that the model can be used to detect the limiting superior and inferior vertebrae in a given FoV with great accuracy. This shows how the model can be applied in the FoV cropping task for subsequent image registration. Overall, the T2w model, while trained on a smaller dataset, performed comparably, if not better than the T1w model. This was due to the employment of transfer learning when training the T2w

Table 3.1: Quantitative metrics used to evaluate the Mask R-CNN model demonstrate robustness in detecting cervical whole vertebra.

Seq.	mAP_{75}	LE [mm] (↓)	IDR (↑)	IRA (↑)	LFA (↑)	Dice (↑)	F1 (↑)
T1w	0.919	2.750	0.960	0.731	0.931	0.858	0.844
T2w	0.867	3.828	0.950	0.844	0.911	0.867	0.863

Table 3.2: Average Precision (AP) values of each vertebra and spinal cord from object detection averaged across IoU thresholds between 0.50 and 0.95 at 0.05 intervals.

Sequence	Cord	C1	C2	C3	C4	C5	C6	C7
T1w	0.897	0.708	0.821	0.785	0.749	0.751	0.674	0.665
T2w	0.911	0.485	0.841	0.805	0.687	0.714	0.792	0.829

model with weights and biases initialized with the pre-trained T1w model.

The AP values of each vertebra are outlined in Table 3.2, where the AP calculated is averaged across IoU thresholds between 0.50 and 0.95 at 0.05 intervals. This results in a stricter metric when observing each class than the mAP_{75} as it uses stricter IoU thresholds (eg. IoU threhsold of 0.95) when calculating. When comparing the AP performance, we see that C2, a distinct landmark vertebra, and the spinal cord are detected at a higher rate compared to other cervical vertebrae. This is expected behaviour similar to manual object detection, where the distinct objects help the model accurately identify and classify with less confusion. In addition, the distinct object C7 with the elongated posterior element was found to also have a higher AP value for the T2w results. In comparison, the C7 AP was the lowest in the T1w model results. This may be attributed to the sequence of the image itself, as the soft tissue surrounding the bony structure in the posterior region of the T2w images are less saturated with information compared to the T1w images. The elongated posterior element is the main differentiation between C7 and the other cervical vertebrae. C1 in the T2w results were the worst performing detection due to the small size (less resolved), resulting in the vertebrae blending in with the surrounding dark contrast soft tissue.

Figure 3.8 a) and b) illustrate the confusion matrices of the T1w and T2w models, respectively, during test set inference after post processing with the graph filter. Both confusion matrices show the models abilities to correctly detect and label the cervical vertebrae. In the T1w model, the spinal cord was always detected correctly, with C1 and C2 either always detected properly, or undetected. This may be attributed to the cropping or slice chosen when processing the dataset resulting in the C1 and C2 vertebrae missing the

posterior element, or perhaps a large enough portion of the C2 vertebrae being cropped out. The graph may have also filtered out the detections from the final output if the vertebrae did not fit into the predetermined rules in constructing the adjacency matrix. The vertebrae C3-C6 were also classified correctly in most cases with some inter-class mis-classifications. Since this is found primarily within the range of C3-C6, the mis-classification may be due to the model predicting the vertebrae with shifted labels due to the vertebrae being very similar in size and shape. For example, the model could have predicted C4-C6 on C3-C5 present in the image. Since the model predictions are still sequential, the graph is unable to determine that there may be a mis-classification error in the inference results, resulting in the small subset of the inter-class mis-classifications. Aside from C4 and C5 in the T2w model, most vertebrae were correctly detected without detection or mis-classification errors.

When observing the "0", or background class, the models tended to detect extra vertebrae false positives (FPs) at the inferior portions of the cervical spine, where vertebrae are very similar in size and shape. These extra vertebrae detections were found to primarily consist of partially cut vertebrae from the data generation of smaller FoVs due to the curvature of the spine relative to the superior-inferior direction, as seen in Figure 3.6 c) and Figure 3.7 c). This was observed most with the partial C7 vertebrae, which correlated to the most extra detections. When these extra detections were processed with the graph, were often correct with respect to the slice, as they followed the sequential hierarchy of the vertebrae, but were undesired when evaluating against the ground truth. These extra detection FPs were considered in the evaluation of the F1 score and IRA metric, leading to a score reduction. As the score threshold increases, the number of FPs from the extra detections decrease in both models. In contrast, a lower threshold increases the amount of detections the model passes to the graph filter, allowing for more detections to be input to the graph for filtering. This allows the graph to correct any mis-classifications that may not be found otherwise. As previously mentioned, the graph is unable to determine a mis-classification error in the inference results if all vertebrae are detected correctly, but are classified with shifted vertebrae labels. While this issue can be mitigated with the lower score threshold, as additional FP detections of the same vertebrae with different vertebral labels allow for the graph to relabel the original predictions, this also increases the chance of allowing extra detections of partially cut vertebrae from being introduced, reducing metric scores. We selected a score threshold of 0.6 to balance the false positive/false negative rate as outlined in Section 3.3.3.

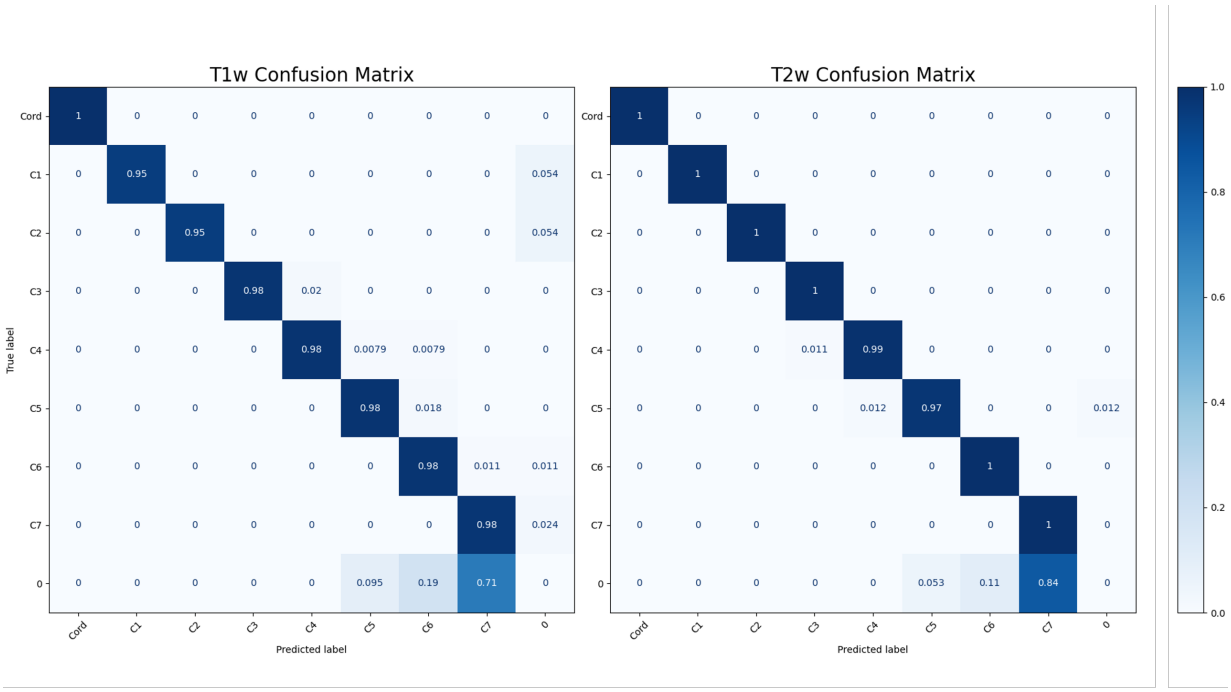


Figure 3.8: Normalized confusion matrices of the (a) T1w and (b) T2w models during test set inference. Class 0 is the background class, showing undetected vertebrae (false negatives) and extra vertebrae detections (false positives).

Table 3.3: Ablation experiment verifying the effectiveness of the graph filter.

Configuration	mAP_{75}	LE [voxels]	IDR	IRA	LFA	Dice	F1
T1w + Graph	0.919	2.750	0.960	0.731	0.931	0.858	0.844
T1w + No Graph	0.920	2.956	0.955	0.569	0.823	0.851	0.831
T2w + Graph	0.867	3.828	0.950	0.844	0.911	0.867	0.863
T2w + No Graph	0.849	4.010	0.940	0.544	0.811	0.866	0.818

3.3.3 Ablation Experiments

Ablation experiments were performed to determine the effect of the graph filter and the confidence score threshold. In addition, examination of the effect of training including or excluding the posterior element of the vertebrae label as also conducted when examining the effects of the score threshold.

Effects of Graph Filter

Table 3.3 shows the comparative metric results with and without the use of the graph filter. The T1w and T2w models with and without the graph show similar performance across most metrics, with the IRA metric being the main differentiation between the use of the graph. The large difference in the IRA metric can be attributed to the graph resolving undetected/misclassified vertebrae and filtering out extra detections, often resulting in perfectly detected vertebrae in the image. However, it would also be expected that the F1 score would also be much lower for the non-graph models. This leads to the conclusion that the graph is fixing a majority of undetected/misclassified vertebrae, but not necessarily removing the extra detections of partial vertebrae, such as C7 as seen in Figure 3.7 c), resulting in similar F1 scores for the models with and without the graph filter.

Effect of Whole Vertebrae Detection

The effect of training the vertebrae detection model using the whole vertebrae, rather than just the vertebral body, was also investigated. Table 3.4 shows the T1w and T2w models trained on whole vertebrae outperform the vertebral body only models in all metrics except for LE and Dice coefficient.

The lower LE and Dice coefficient of the vertebral body only models can be attributed to the smaller localized bounding boxes and masks. The bounding boxes surrounding the

Table 3.4: Ablation experiment verifying the effectiveness of training to detect the whole vertebrae rather than only the vertebral body.

Configuration	mAP_{75}	LE [mm]	IDR	IRA	LFA	Dice	F1
T1w + Whole Vert	0.919	2.750	0.960	0.731	0.931	0.858	0.844
T1w + Vert Body	0.778	1.654	0.886	0.638	0.815	0.963	0.822
T2w + Whole Vert	0.867	3.828	0.950	0.844	0.911	0.867	0.863
T2w + Vert Body	0.711	1.725	0.864	0.467	0.878	0.967	0.755

whole vertebrae generate centroids that are highly sensitive to the orientation and shape of the vertebrae. In addition, the larger size of the bounding box compared to the smaller region detected of the vertebral body results in the larger error seen from centroid LE. The lower Dice coefficient is caused primarily by the segmentation of the posterior element, which is often poorer due to the surrounding soft tissue causing confusion in inference. The vertebral body is easily segmented due to it's distinct shape and size, whereas the posterior element can vary greatly in size and shape depending on the sagittal slice used. This leads to the Dice coefficient being much greater in the vertebral body only model as it has a much easier segmentation task.

The relatively similar IRA and F1 score metrics between the T1w vertebral body and whole vertebrae models in comparison to the larger difference between the T2w models can be attributed to the extra detections from partially cut vertebrae. As seen in Figure 3.9 a), large portions of the posterior element still present that result in vertebrae detections. In comparison, the vertebral body T1w model does not recognize the posterior element, resulting in no extra detection as seen in Figure 3.9 b).

Effect of Score Threshold

Results for the ablation experiment using score thresholds ranging from 0.0-0.9 in 0.1 increments on the whole vertebrae T1w and T2w models are outlined in Table 3.5 and Table 3.6, respectively. Observing the experiment results, there is no ideal score threshold based on best overall performance in all metrics. A lower score threshold tends to benefit a higher mAP_{75}, IDR, and LFA. In comparison, a higher score threshold tends to benefit the LE, IRA, Dice, and F1 score. This is primarily due to the lower score threshold increasing the amount of total detections that are then processed with the graph. As previously mentioned, the higher number of detections tends to result in extra detections of partially cut vertebrae in the images, resulting in a lower IRA and F1 score. This can also lead to higher errors with the LE and Dice as the extra detections are penalized. In contrast,

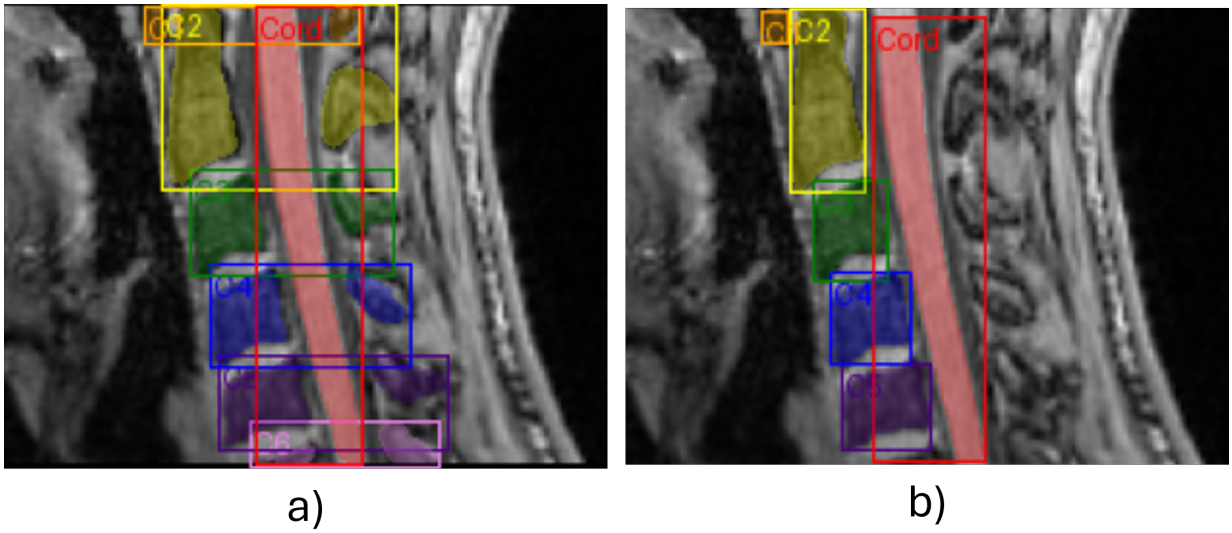


Figure 3.9: Inference results of T1w whole vertebrae detection vs. vertebral body only detection of the same subject. a) T1w whole vertebrae detection of C1-C5, with extra detection of C6. b) T1w vertebral body detection of C1-C5, without extra detection.

the higher threshold limits detections, potentially resulting in undetected vertebrae. A balanced score threshold is then required to ensure enough predictions are generated to detect all vertebrae, without too many extra detections. This resulted in the chosen score threshold of 0.6 for both the T1w and T2w models, where IRA is maximized, with minimal decreases or increases in other metrics.

3.3.4 State-of-the-Art Detection Model Comparison

Model comparisons were conducted to evaluate the performance of the proposed Mask R-CNN for vertebrae detection compared to other state-of-the-art methods. Two comparisons were made: 1) evaluating performance against the Faster R-CNN architecture for a similar task, and 2) evaluating performance against other methods. The models that we compare to are not trained on the same data, nor only on the cervical spine, but do give some reference to the performance of our vertebrae and spinal cord detection model.

As shown in Table 3.7, we compare the results of our cervical vertebrae and spinal cord detector to Faster R-CNN implementations found in literature detecting vertebrae in MRI ([50, 39, 40]). These Faster R-CNN models were trained on lumbar/thoracic spine MRI

Table 3.5: Ablation experiments to determine score threshold effect on graph filtered inference results using T1w whole vertebrae model.

Score Threshold	mAP_{75}	LE [mm]	IDR	IRA	LFA	Dice	F1
0.0	**0.926**	2.628	**0.970**	0.300	**0.954**	0.846	0.810
0.1	0.923	2.848	0.967	0.338	0.946	0.846	0.815
0.2	0.880	3.982	0.950	0.246	0.877	0.837	0.799
0.3	0.925	2.649	0.969	0.500	0.946	0.852	0.828
0.4	0.925	2.649	0.969	0.569	0.946	0.854	0.833
0.5	0.922	2.643	0.966	0.631	0.931	0.856	0.837
0.6	0.919	2.750	0.960	**0.731**	0.931	0.858	0.844
0.7	0.900	**2.592**	0.938	0.715	0.908	**0.859**	**0.845**
0.8	0.861	2.718	0.885	0.654	0.769	0.857	0.820
0.9	0.691	3.397	0.707	0.300	0.554	0.849	0.713

Table 3.6: Ablation experiments to determine score threshold effect on graph filtered inference results using T2w whole vertebrae model.

Score Threshold	mAP_{75}	LE [mm]	IDR	IRA	LFA	Dice	F1
0.0	0.869	3.741	**0.956**	0.267	**0.933**	0.853	0.800
0.1	0.869	3.741	0.954	0.289	0.922	0.854	0.813
0.2	**0.874**	3.742	**0.956**	0.400	0.922	0.855	0.828
0.3	0.869	3.898	0.952	0.578	0.911	0.859	0.842
0.4	0.869	3.898	0.952	0.689	0.911	0.862	0.851
0.5	0.869	3.894	0.952	0.778	0.911	0.865	0.860
0.6	0.867	3.828	0.950	**0.844**	0.911	0.867	**0.863**
0.7	0.848	3.797	0.922	0.822	0.889	0.866	0.846
0.8	0.782	3.967	0.820	0.511	0.789	**0.870**	0.791
0.9	0.563	**2.897**	0.568	0.089	0.544	0.862	0.589

Table 3.7: Comparison of cervical spine Mask R-CNN vertebrae and spinal cord detector to Faster R-CNN detection results for mixed T1w/T2w scans of lumbar/thoracic vertebrae.

Work	Sequence	mAP_{75}	IDR	IRA	LE [mm]
Ours	T1w	0.919	0.960	0.731	2.750
Ours	T2w	0.867	0.950	0.844	3.828
Zhao *et al.* [50]	T1w/T2w	0.829	0.884	0.831	7.124
Zeng *et al.* [39]	T1w/T2w	0.848	0.869	0.750	-
Zhao *et al.* [40]	T1w/T2w	-	0.884	-	7.124

in T1w and T2w, whereas our models are trained on the T1w and T2w cervical spin MRI separately. While the regions of the spine are different, the lumbar/thoracic data is similar such that the lumbar spine has the sacrum as a distinct landmark vertebra, similar to C2. We find that our Mask R-CNN model with the post process graph filter outperforms the Faster R-CNN models in the mAP_{75}, IDR, LE metrics. This can be attributed to the Mask R-CNN's architecture being based on the Faster R-CNN architecture with additional improvements. In addition, the use of the graph filter helps correct any incorrect classifications to ensure sequential hierarchy is enforced. However, the models perform similarly to the Faster R-CNN models when observing the IRA metric. This is attributed to the extra detections of partial vertebrae reducing the metric score.

When evaluating the Mask R-CNN vertebrae and spinal cord detection model performance against other state-of-the-art vertebrae detection models (Table 3.8), we find that comparable performance across all metrics except for IRA. It is important to highlight that the state-of-the-art deep learning models being used for comparison are not trained on the same data, with their results used only as a reference.

The T1w model mAP_{75} was comparable to the performance reported by Zhao *et al.* and Zeng *et al.*; however, the T2w model performed relatively worse. While this means the model was less accurate and precise at the 75% IoU threshold, the model still performs well as a vertebrae detector. This lower mAP_{75} may be attributed to the smaller dataset size used to train the our T2w model compared to others. Both models were able to detect present vertebrae with a similar rate to the state-of-the-art models as observed by the comparable IDR metrics. In contrast, the T1w and T2w models had much lower IRA. This is primarily due to the extra detections caused by partial vertebrae in the image slices as previous outlined. LE performance was also comparable to other models only computing the bounding box of the vertebral body.

Table 3.8: Comparison of cervical spine MRI Mask R-CNN vertebrae and spinal cord detector to state-of-the-art deep learning vertebrae MRI detection models.

Work	Sequence	Region	mAP_{75}	IDR	IRA	LE
Ours	T1w	Cervical	0.919	0.960	0.731	2.750
Ours	T2w	Cervical	0.867	0.950	0.844	3.828
Forsber *et al.* [41]	T1w	Cervical	-	0.960	-	1.180
Forsber *et al.* [41]	T2w	Cervical	-	0.960	-	1.240
Zhao *et al.* [50]	T1w/T2w	Lumbar/Thoracic	0.964	0.963	0.933	5.199
Chang *et al.* [54]	T1w/T2w	Lumbar/Thoracic	-	0.893	-	-
Zeng *et al.* [39]	T1w/T2w	Lumbar/Thoracic	0.973	0.969	0.973	-
Zhao *et al.* [40]	T1w/T2w	Lumbar/Thoracic	-	0.974	-	4.742
Windsor *et al.* [51]	T1w/T2w	Whole Spine	-	0.965	-	2.400

3.4 Conclusions

This investigation demonstrated that a Mask R-CNN object detection model was capable of accurately localizing and labelling individual vertebra and the spinal cord within the cervical spine in 2D anatomical MRI images. In addition, it was shown that leveraging the sequential hierarchy of the vertebral levels using a graph can increase detection performance. Overall, the Mask R-CNN model shows state-of-the-art performance, robust to varying cervical spine FoVs. The model outputs bounding boxes and masks that can be used for accurate FoV cropping and initial affine registration prior to deformable registration, towards the goal of automating multi-modal image registration of the cervical spine.

Chapter 4

Application of Vertebral Detection Models for Multi-Modal Medical Deformable Image Registration

This chapter discusses the application of the Mask R-CNN cervical spine detection and labelling model presented in Chapter 3 in a multi-modal and multi-sequence deformable image registration pipeline.

4.1 Introduction

Multi-modal deformable image registration for spinal applications typically requires the fixed and moving image pairs to have the same spatial features, or vertebrae field of view (FoV). Use of images with differing spatial features will often result in deformable registration failure. Chapter 3 outlines the development of a deep learning cervical spine detection model for MRI as a first step towards improving automated deformable image registration. With the use of the detection model, the vertebrae can be identified in both the fixed and moving images to assist with image FoV cropping and affine transformation, prior to application of deformable image registration techniques.

In this chapter, we propose the development of an automated multi-modal deformable image registration pipeline for clinical data of the spine consisting of four components: 1) cervical vertebrae detector for MRI and CT scans, 2) a FoV cropping component using the detected vertebrae, and 3) an affine registration component, utilizing scaling, rotation,

and translation (no shearing) to seed initial alignment of the FoV cropped images, and 4) a deformable image registration module. A multi-sequence image registration pipeline was first developed to demonstrate the vertebrae detection and FoV cropping with T1w and T2w MRI. The pipeline was then adopted for a multi-modal deformable image registration pipeline between MRI and CT images.

4.2 Methods

To evaluate the use of the previously proposed Mask R-CNN cervical spine detection model to automate multi-modal spine image registration, we first developed a multi-sequence spine image registration pipeline. Specifically, the pipeline for multi-sequence cervical spine MRI registration consisted of sequential integration of the trained Mask R-CNN cervical spine detection model, along with a FoV cropping algorithm, followed by affine registration algorithm, and finally the use of a continuous optimization deformable registration [60]. This deformable registration pipeline was initially examined using an intra-subject, multi-sequence dataset of T1w and T2w images. Additionally, extension to multi-modal image registration of the spine, including both MRI and CT, was performed with the addition of a CT vertebrae detection model is outlined in Section 4.4.

The evaluation of the pipeline was performed on the T1w and T2w scans of three subject from Chapter 3, unseen by the model during training, validation, and testing. Figure 4.1 outlines the multi-sequence image registration pipeline.

4.2.1 Multi-Sequence Deformable Registration Pipeline

Three subjects from the Spine Generic dataset were chosen for intra-subject, multi-sequence deformable registration, previously unseen by the detection model during training, validation, and testing. The MRI scans were processed utilizing the methods outlined in Chapter 3.2.1 to generate ground truth segmentations of the vertebral level on the spinal cord for registration evaluation. The T2w image of a subject were utilized as the fixed, stationary image during registration. In contrast, the T1w image were used as the moving image, being warped into spatial alignment with the fixed image. In addition, the fixed T1w images contained smaller, varying FoVs, whereas the moving T2w images contained a larger FoV containing all cervical vertebrae (C1-C7). This corresponds to having a larger FoV CT scan as the moving image being registered to a smaller FoV MRI scan. Eight fixed T1w scans with differing vertebrae FoVs of the same subject were used including: C1-C6,

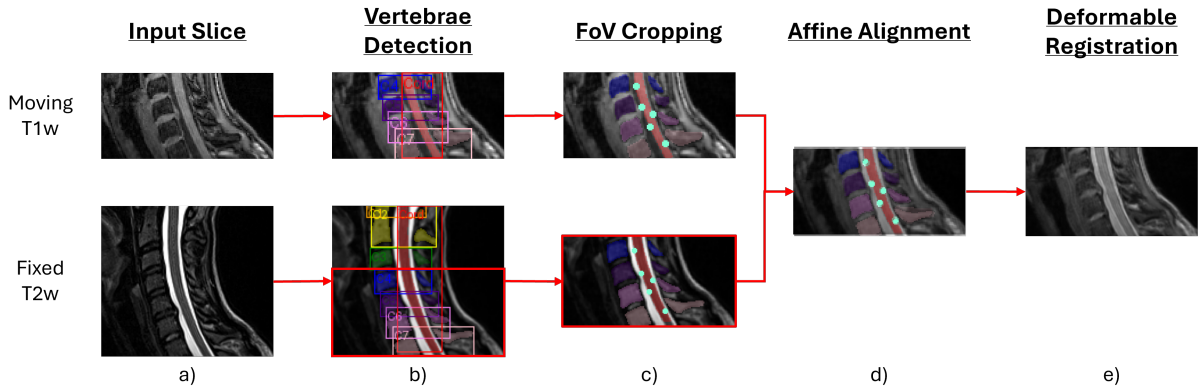


Figure 4.1: Intra-subject, multi-sequence image registration pipeline for T1w to T2w deformable registration. a) Input mid-sagittal slice. b) Feature detection using Mask R-CNN cervical vertebrae and spinal cord MRI detection model. c) Cropping fixed or moving image to limiting FoV. d) Initial affine registration using bounding box centroids of moving image overlaid on fixed image. e) Continuous optimization-based deformable image registration of moving image overlaid on fixed image.

C1-C5, C1-C4, C3-C7, C3-C6, C3-C5, C4-C7, C4-C6. Figure 4.1 a) shows an example of the fixed and moving image pair data of a single subject input to the multi-sequence spine registration pipeline. The fixed T1w image contains four vertebrae, compared to the moving T1w image containing seven vertebrae.

Cervical Vertebrae and Spinal Cord Detection Inference

Cervical spine detection was performed using the deep learning Mask R-CNN model developed in Chapter 3. Inference was performed on the selected subject image pairs, with the fixed and moving 3D volumes used as inputs to the registration pipeline to mimic clinical usage of 3D deformable registration. The mid-sagittal slice of the volumes were then extracted, with image resizing transformations applied prior to model inference. These transformations were performed to ensure input data was sized similarly to data used during training. This includes performing min-max normalization, resampling to a voxel size of 1 mm^3, resizing the smaller dimension of the images to a max size of 225 voxels, and zero padding to a resulting image size of 250x250 voxels. The pre-processed slices were then input to the model for inference as seen in Figure 4.1 b). The moving T1w image was found to contain C4-C7, whereas the fixed T2w image contains C1-C7. The spinal cord was also detected in both images. Inference outputs were transformed back to the original

resampled slice, then resampled back to the original grid size for subsequent FoV detection and cropping.

Field of View Detection and Cropping

Using the inference outputs generated of the fixed T2w slice and moving T1w slice, a FoV detection and cropping algorithm was developed. First, the limiting (smaller) FoV between the fixed and moving images was determined. This was performed by observing the superior and inferior-most inference output labels generated in the fixed and moving images as the vertebrae follow a sequential hierarchy. In this case, the moving T1w image contained C4-C7, therefore requiring the fixed T2w image to be cropped to contain the same vertebrae FoV as seen in Figure 4.1 c). With the limiting FoV vertebrae and associated T2w image known, cropping was performed on the larger FoV image in the SI axis. The superior-most C4 coordinate and inferior-most C7 coordinate obtained from their respective bounding boxes were used to crop the fixed image and masks.

The bounding box of the spinal cord in the fixed image also needed to be recalculated to ensure the spinal cord region overlaps with the region in the moving image. The spinal cord's bounding box in the fixed image was regenerated based on the resultant cropped spinal cord segmentation (such that the bounding box region overlapped with moving image). As the spinal cord is cropped in the superior-inferior (SI) axis, the subsequent anterior-posterior (AP) length of the bounding box was larger than required due to the curvature of the spine as seen in Figure 4.2 a). This led to the requirement to completely resize the bounding box based on the cropped spinal cord segmentation, as seen in Figure 4.2 b), to ensure an accurate spinal cord centroid for affine registration between the fixed and moving image.

Affine Registration

To perform affine registration, transform model parameters are first calculated. This includes first extracting 3D bounding box centroid coordinates from the processed and FoV cropped inference output fixed and moving images. The first two coordinates of the bounding box centroids are obtained from the bounding box prediction on the 2D slice. The third coordinate is known due to the use of the mid-sagittal slice of the input volume. These centroids were then used to generate an affine transformation matrix used to register the moving volume to the fixed volume.

The affine transformation matrix, $\mathbf{T}$, can be calculated using Equation 4.1 and a least squares method to minimize total error of matching vertebrae centroid distances. Here,

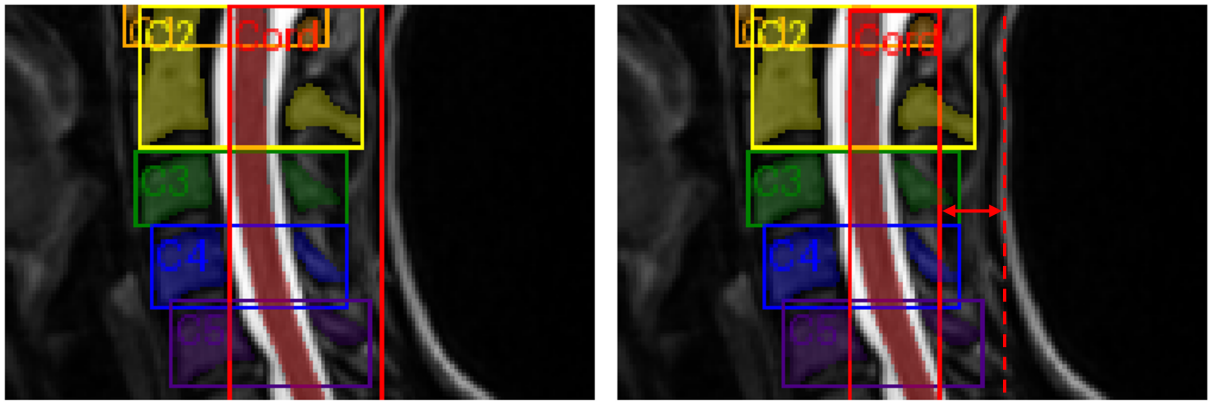


Figure 4.2: Regeneration of spinal cord bounding box due to cropping in the SI axis. a) Cropped image (originally C1-C7) in SI axis results in an oversized spinal cord bounding box in the AP axis. b) Regenerated spinal cord bounding box based on the cropped spinal cord segmentation results in a shrunken bounding box in AP axis.

$\mathbf{y}$ is the transformed output and $\mathbf{x}$ is the input. $\mathbf{y}$ is the centroids of the T2w fixed image, whereas $\mathbf{x}$ is the centroids of the T1w moving image. $\mathbf{T}$ is a 4×4 homogeneous transformation matrix consisting of a rotation matrix, $\mathbf{R}$, and translation vector, $\mathbf{b}$. Note that Equation 4.1 is a rewritten form of Equation 2.1 for 3D affine transformation.

$$\begin{gathered}\mathbf{y} = \mathbf{T}\mathbf{x} \\ \begin{bmatrix} y_x \\ y_y \\ y_z \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & b_x \\ R_{21} & R_{22} & R_{23} & b_y \\ R_{31} & R_{23} & R_{33} & b_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_x \\ x_y \\ x_z \\ 1 \end{bmatrix}\end{gathered} \tag{4.1}$$

With the affine transformation matrix generated from the 3D centroids, it can then be applied to the moving image to register the image to roughly the same anatomical space as the fixed image as seen in Figure 4.1 d).

Due to the sequential nature of the vertebrae centroids in a 2D slice, the vertebrae centroids from the bounding box inference create a line in the SI direction as seen in Figure 4.3 a). When solving for the affine transformation matrix, the least squares method used may mirror and rotate the image, as seen in Figure 4.3 b), as the centroids are still

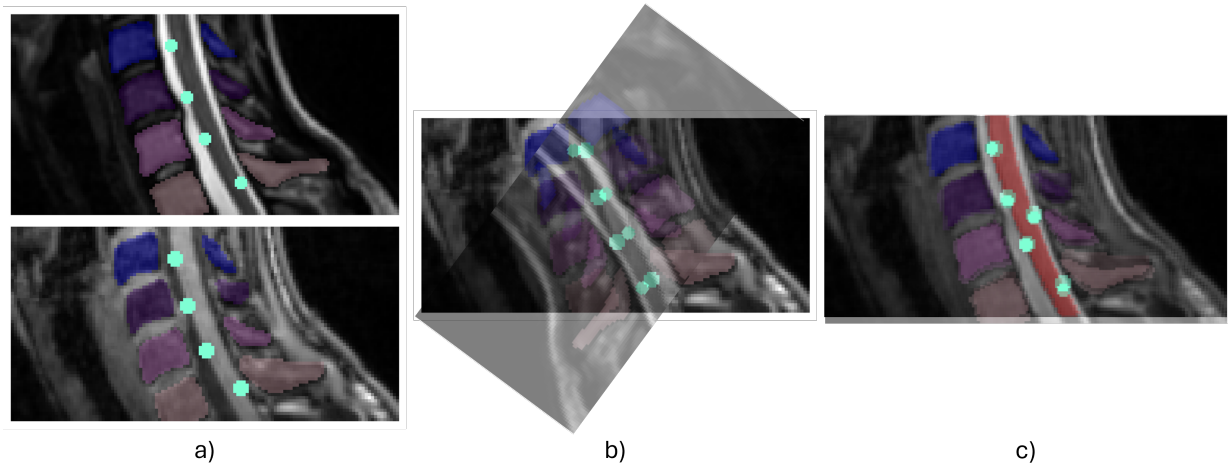


Figure 4.3: Moving image orientation potentially mirroring and rotating when obtaining affine transformation matrix. a) Cropped fixed and moving images with vertebrae centroids. b) Mirrored and rotated moving image from applied affine transformation without spinal cord centroid. c) Affine transformation with vertebrae and spinal cord centroids enforce image orientation.

able to be matched accordingly. Figure 4.3 c) shows how the addition of the spinal cord centroid can enforce the image orientation during initial affine registration.

Deformable Image Registration

Deformable image registration of the affine aligned cervical spine images was performed using a continuous optimization-based registration algorithm [60].

The algorithm implements a continuous optimization of the objective function (Equation 4.2 to determine the optimal displacement field or mapping function, ϕ^*, similar to Equation 2.2. Here, we find the optimal displacement field using a gradient descent method described in Equation 4.3, where L is the loss function described by the contents of the max function outlined in Equation 4.2, η is the learning rate, and ∇_ϕ is the Jacobian operator.

$$\phi^* = \max_{\phi}(\mathcal{S}(f, m \circ \phi) + \mathcal{R}(\phi)) \tag{4.2}$$

$$\phi^t = \phi^{t-1} - \eta \nabla_\phi L \tag{4.3}$$

To compare similarity, or image alignment between the fixed and moving images, Equation 4.4 is utilized. Here, N is the size of the image in voxels, F_f is the modality independent neighbourhood descriptor (MIND) feature of the fixed image, and F_m is the MIND feature of the moving image.

$$\mathcal{S} = \frac{1}{N}(F_f - F_m)^2 \tag{4.4}$$

The regularization function, Equation 4.5, is the mean gradient of the displacement field.

$$\mathcal{R} = \frac{-1}{N}|\nabla D| \tag{4.5}$$

In addition, the algorithm uses an ADAMs optimizer to augment the gradients with momentum, in addition to a piece-wise decreasing learning rate, for the continuous optimization of the optimal displacement field. Once the optimization has been performed to produce an optimal displacement field, it is applied to the moving image, resulting in a registered image pair as seen in Figure 4.1 e).

4.2.2 Pipeline Evaluation Criteria

Experiments were performed to validate the multi-sequence deformable image registration pipeline and the use of the cervical spine detection model.

Qualitative Performance

Qualitative evaluation of the multi-sequence deformable image registration was conducted to demonstrate the pipeline's ability to detect the cervical vertebrae and spinal cord, apply a field of view cropping, perform initial affine registration, and finally deformable registration. Specifically, T1w moving images, with smaller and varying FoVs, were registered to T2w fixed images with a fixed larger FoV. Here, the warped images, segmentations from inference, and ground truth segmentations of the vertebral levels on the spinal cord can be visualized to evaluate registration performance.

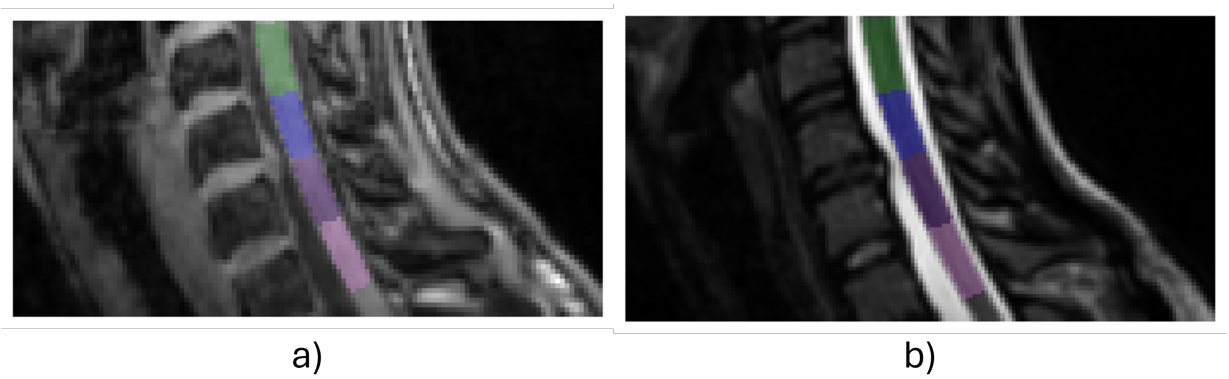


Figure 4.4: SCT segmentation of the vertebral level on the spinal cord used for quantitative evaluation of the multi-sequence deformable registration pipeline. a) T1w moving image. b) T2w cropped fixed image.

Quantitative Performance

Quantitative evaluation of the multi-sequence image registration pipeline was conducted using the Dice coefficient metric. This metric is a common segmentation metric, describing the overlap of the fixed and moving segmentations after registration. We calculate the Dice metric between the fixed and moving ground truth vertebral level segmentations on the spinal cord in the 3D MRI volume, and 2D mid-sagittal slice, as seen in Figure 4.4 a) and Figure 4.4 b). This segmentation was used for the Dice coefficient calculation due to it having information on the vertebral levels, while also being the most accurately generated segmentation from SCT during the data generation process, allowing for 3D deformable registration evaluation. This ensured that any incorrect vertebrae detections or registration errors would result in a low Dice coefficient due to the mismatch of the vertebral levels on the spinal cord.

4.3 Results and Discussion

The multi-sequence deformable image registration pipeline was successfully developed for use with the cervical spine. Data from three subjects was used to perform intra-subject image registration between a fixed T2w and moving T1w image.

Table 4.1: Quantitative Dice coefficient metric used to evaluate the deformable registration performance from the multi-sequence spine image registration pipeline.

Input	Subject	Average Dice Coeff.	Total Average Dice Coeff.
Volume	sub-beijingPrisma03	0.851	0.828
	sub-nwu02	0.790	
	sub-tokyoSkyra01	0.842	
Slice	sub-beijingPrisma03	0.742	0.744
	sub-nwu02	0.731	
	sub-tokyoSkyra01	0.761	

4.3.1 Qualitative Evaluation of the Multi-Sequence Deformable Image Registration Pipeline

Figure 4.5 show examples of the deformable registration results between a fixed T2w image, and moving T1w image of two subject. The pipeline demonstrated the ability to detect cervical vertebrae and the spinal cord, crop differing FoVs, and apply subsequent affine and deformable registration. Figure 4.5 a) depicts the moving T1w image superimposed on the FoV cropped T2w image and it's predicted vertebrae segmentations based on the location in anatomical space. Figure 4.5 b) shows the the moving T1w affine aligned to the fixed T2w image using the bounding box centroids from inference. Lastly, the continuous optimization-based deformable registration was applied to the moving T1w image in Figure 4.5 c).

4.3.2 Quantitative Evaluation of the Multi-Sequence Deformable Image Registration Pipeline

Table 4.1 outlines the average dice coefficient for the intra-subject deformable registration of the fixed and moving volumes, and the slices used for inference. The registration performed exceptionally well, with an overall Dice coefficient of 0.828 for the volume registration, and 0.744 for the slice registration. The lower Dice coefficient from the slice registration is expected due to the registration pipeline being designed for volume registration. Since the mid-sagittal slices of the volume are used for registration, the fixed and moving image may not have the same anatomical slice, resulting in larger registration errors. In comparison, the 3D volume is able to utilize the third ML dimension and perform registration with greater accuracy.

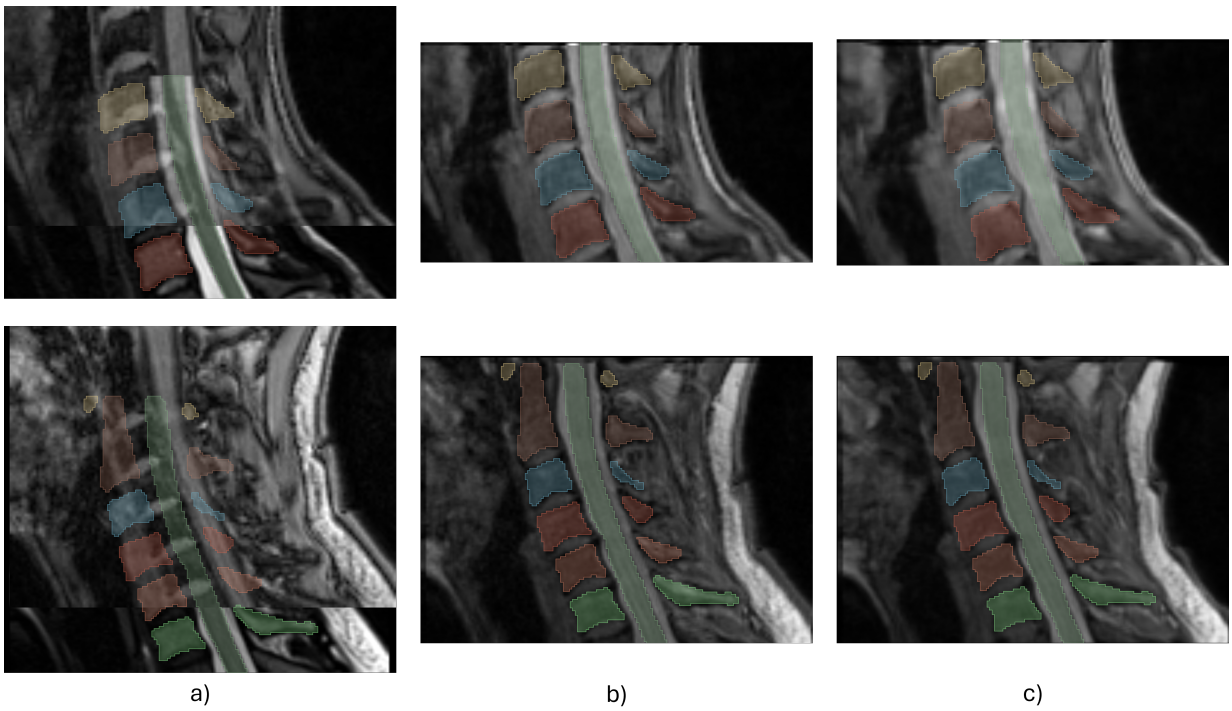


Figure 4.5: Spine Generic T1w to T2w deformable registration results of two subjects. The moving T1w image is superimposed on the cropped fixed T2w image and T2w inference vertebrae segmentation. a) Moving T1w image and fixed T2w image in original physical space. b) Affine aligned moving T1w to T2w image using inference bounding box centroids. c) Deformable registration applied to moving T1w image.

4.4 Application to Clinical Data

Spine treatment planning (surgery, radiotherapy, etc.) commonly involves the use of multiple imaging modalities. CT is used to image the spine because it has excellent bone contrast, visualizing disease or trauma to the bone well. MRI is used because it has excellent soft tissue contrast, which is essential for visualizing the neural elements, and any soft tissue disease, such as cancer.

Stereotactic body radiotherapy (SBRT) is commonly used to treat tumours in the bony spine. SBRT involves the use of multiple beams of radiation from multiple angles, delivering a high dose of vocal radiation to the tumour. Treatment planning for SBRT involves the use of both CT and MRI. The CT is used to visualize bony tumour involvement, and MRI is used to visualize the neural elements of the spine and any soft tissue tumours. SBRT planning involves contouring radiotherapy targets that include the tumour, as well as regions of the body that should be spared irradiating, such as the spinal cord and nerve roots. In addition, SBRT requires registration of the MRI and CT prior to treatment planning. Currently, registration is performed manually by manual translation and rotation for the vertebrae being ablated, taking considerable manual effort, potentially requiring adjustment if multiple vertebrae are ablated.

To automate and improve spine registration, the trained Mask R-CNN model was applied to a multi-modal (CT and MRI) deformable image registration pipeline, adopting from the multi-sequence registration pipeline. This required the addition of a CT vertebrae detection branch. The multi-modal pipeline was evaluated on a clinical dataset containing both CT and MRI scans of the cervical spine, with a focus on the performance on the Mask R-CNN vertebrae and spinal cord detection model.

4.4.1 Spine Metastasis Dataset

The spine metastasis dataset (SpineMets), provided by collaborators at the Sunnybrook Research Institute, is a prospectively maintained database encompassing all patients with spinal metastases treated with spine SBRT at a single institution from January 1, 2008, to December 31, 2022. During this period, a total of 977 patients with 2,364 spinal segments received treatment. The database includes comprehensive baseline clinical and tumour characteristics, as well as outcome measures, such as fracture development and overall survival rates. For each patient, the database contains treatment planning CT and MRI scans, including anatomical T1w and T2w imaging. Patients were de-identified and converted from DICOM to NIfTI format for research use. All CT and MRI scans had varying

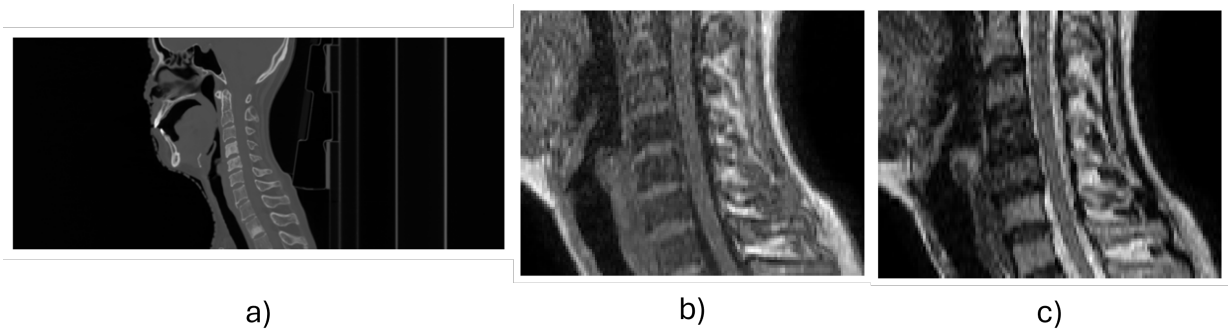


Figure 4.6: A single subject from the SpineMets clinical dataset collected by Sunnybrook Research Institute. Pathology observed in the C4 vertebrae. a) CT scan containing C1-T3 vertebrae. b) T1w scan containing C2-C7. c) T2w scan containing C2-C7.

scanning parameters depending on the scan and machine used.

Seven subjects within the SpineMets dataset were selected for multi-modal image registration, ensuring the MRI scans contained only cervical spine data. Figure 4.6 a), Figure 4.6 b), and Figure 4.6 c) illustrate an example of the CT, T1w MRI, and T2w MRI scan of a selected subject, respectively. It can be observed that the CT scans contain the C1-T3 vertebrae, whereas the T1w and T2w MRIs have a smaller FoV, containing C2-C7. The CT scans were used as the fixed images and MRI as the moving images. Selected CT scans used were captured on a Phillips Brilliance Big Bore CT machine, with images having a non-isotropic voxel size of approximately 1.2mm, 1.2mm, 1.0mm. Selected T1w MRI were imaged on a 1.5T Phillips Ingenia and have non-isotropic voxel sizes (RAS) of approximately 0.8mm, 0.8mm, 1.0mm or 0.85mm, 0.85mm, 2.0mm, due to differing scanning parameters. Selected T1w MRI have non-isotropic voxel sizes (RAS) of approximately 0.8mm, 0.8mm, 1.0mm or 0.85mm, 0.85mm, 2.0mm, due to differing scanning parameters. The MRI scans were also observed to be lower quality compared to the Spine Generic data used to train the Mask R-CNN model. This can be attributed to the larger voxel size in the SI axis compared to the Spine Generic dataset, leading to a lower resolution observed in the sagittal slices.

4.4.2 Multi-Modal Deformable Image Registration Pipeline for Clinical Dataset

To perform multi-modal deformable image registration of the spine between MRI and CT scans, a modified clinical pipeline was developed, adopting from the multi-sequence registration pipeline. Figure 4.7 outlines the multi-modal pipeline for CT to MRI deformable registration. This pipeline adopted from the multi-sequence pipeline, but with modifications to the affine registration and FoV cropping algorithm. In addition, it utilized VertDetect [45] to process CT scans and detect present vertebrae. VertDetect was trained on the VerSe 2019 and 2020 challenge datasets and has been previously applied on the SpineMets dataset for clinical annotations. First, vertebrae detection is performed on the MRI and CT image pair, using the Mask R-CNN vertebrae and spinal cord detection model and VertDetect, respectively. Both detectors utilized the 3D volumetric scans as inputs, with VertDetect detecting vertebrae and bounding boxes of vertebrae in 3D. In comparison, the Mask R-CNN vertebrae and spinal cord detector can only be applied to a 2D slice, resulting in the default use of the volumetric mid-sagittal slice. Manual selection of an optimal mid-sagittal slice can also be input for the vertebrae and spinal cord detection. Then, affine registration, without shearing, is applied using the detected vertebrae centroids to align the moving MRI to the fixed CT scan in anatomical space. Next, the larger FoV is cropped to the same volume size and space as the affine registered MRI scan. This was performed as the MRI and CT contain vastly different spatial information, compared to the generated Spine Generic dataset, which only differed in the SI axis. In addition, by applying the affine registration without shear using the detected vertebrae centroids, undetected/mislabelled vertebrae could still be properly registered. Lastly, deformable registration was applied utilizing a continuous optimization-based deformable registration algorithm [60].

4.4.3 Vertebrae and Spinal Cord Detection

The MRI and CT scans of seven subjects from the SpineMets dataset, were processed using VertDetect and the Mask R-CNN cervical vertebrae and spinal cord detector. VertDetect was found to localize and label all vertebrae in the selected subjects. Detailed results and analysis of performance of VertDetect are excluded from the contents of the thesis as the focus was on the evaluation of the Mask R-CNN detection model applied to the clinical SpineMets MRI data.

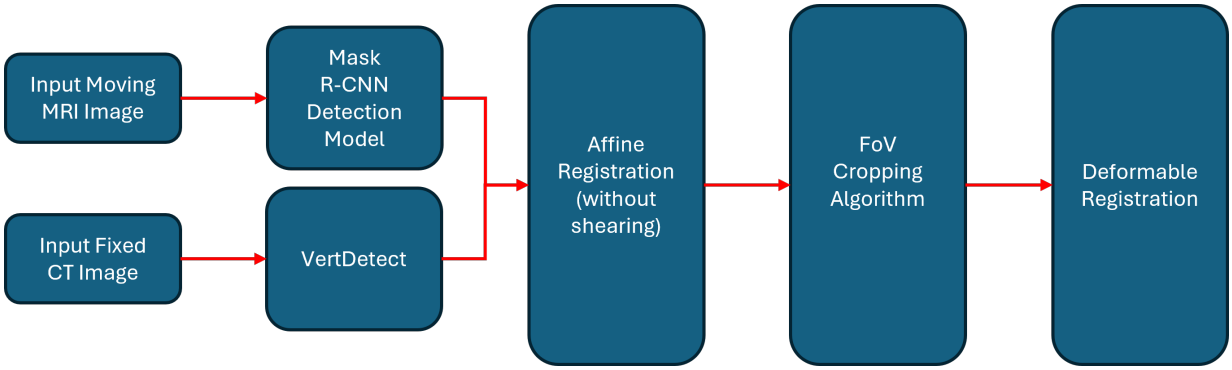


Figure 4.7: Multi-modal spine image registration pipeline for CT to MRI (T1w or T2w) deformable registration. This pipeline adopts from the multi-sequence spine image registration pipeline, with the addition of the VertDetect head for CT vertebrae detection and modifications for the affine registration and FoV cropping algorithm.

Cervical Vertebrae and Spinal Cord Detection Inference on Clinical Data

The Mask R-CNN cervical vertebrae and spinal cord detection model was applied to the SpineMets clinical dataset. The mid-sagittal slice of the input MRI volume was first utilized to evaluate the performance in a fully automated scenario. Figure 4.8 and Figure 4.9 show example inference results on T1w and T2w mid-sagittal slices of four subjects: Subject 1, Subject 2, Subject 3, and Subject 4. T1w and T2w inference outputs of the remaining subjects can be found in Appendix B.

The T1w model was found to perform poorly, with no fully correct detections and many images not having any detections at all. In addition, the model had difficulties detecting the spinal cord, which is one of the most distinct features in the scan. Using the identification rate (IDR) and image recognition accuracy (IRA) metrics from Chapter 3, we observe an IDR of 0.16 and IRA of 0. The poor performance may be attributed to the difference in scanning parameters for the SpineMets T1w images in comparison to the Spine Generic data used for model training. Specifically, the contrast of anatomy in the SpineMets T1w images differs greatly compared to the Spine Generic data. This can be observed in the similar bone to soft tissue contrast and spinal cord to soft tissue contrast. In addition, clinical T1w images are lower quality and contain pathology that the model may not have generalized to when trained on the Spine Generic data, with those subjects being considered healthy and absent of spinal pathology. Subject 1 (Figure 4.8 a)) has a FoV that is more superior than any of the data used to train the model. In addition, the C1 and C2 vertebrae observed indicate this slice is not mid-sagittal to the anatomy, resulting

in the shorter, cut-off C2 process. An incorrect detection of C6 was also predicted. Subject 2 (Figure 4.8 b)) had the best detection for the T1w images. The model was successful in detecting C3-C7 and the spinal cord correctly, while only missing the C2 vertebrae. This can be attributed to the clean input slice with clear vertebral bodies and posterior elements. Subject 3 (Figure 4.8 c)) showed the model failing to detect any vertebrae. The image quality for this subject was poor, with kyphosis in the cervical spine, rather than the lordosis in a healthy spine that the model is exposed to in training. Lastly, Subject 4 (Figure 4.8 d)) only had the model detect C2. Overall, no T1w image had all vertebrae correctly detected.

In comparison, the T2w model performed extremely well, with all images containing correct detections, resulting in an IDR of 0.80 and IRA of 0.14. While the image quality is degraded compared to the Spine Generic data used to train the model, the general image features and feature contrast were similar. Subject 1 (Figure 4.9 a)) had correct detections of C2-C5 and the spinal cord, with a missing C1 detection. This may be attributed to the input slice not being mid-sagittal to the anatomy as observed by the C1 and C2 vertebrae. Subject 2 (Figure 4.9 b)) had the model correctly detecting and labelling all present vertebrae and the spinal cord. Subject 3 (Figure 4.9 c)) correctly detected C1-C4 and the spinal cord, with C5 being undetected and C6 being mislabelled as C5. Due to the kyphosis of the cervical spine, and the observed image artifacts, the model performed extremely well in the detection of the vertebrae compared to the T1w model's performance on the same subject. Lastly, Subject 4 (Figure 4.9 d)) had detections of C1-C4 and the spinal cord, but missed the C5 vertebrae.

Additionally, Figure 4.10 shows the VertDetect inference results of Subjects 1. When observing the input CT scan in Figure 4.10 a), we observe a larger FoV, where the image contains the entire cervical spine and part of the thoracic spine. Figure 4.10 b) shows the VertDetect inference results correctly segmenting and labelling present vertebrae in 3D.

Effect of Manual Selection of the Mid-Sagittal Slice on Clinical Data Inference Due to the mixed performance of the Mask R-CNN model on the mid-sagittal slice of the input clinical data volumes, we experimented with manual selection of the mid-sagittal slice for inference. This was done to select slices that were more anatomically correct as the mid-sagittal plane, mimicking the Spine Generic data used to train the model. Figure 4.11 and Figure 4.12 show the inference outputs on the T1w and T2w optimal mid-sagittal slice of Subjects 1-4, respectively. T1w and T2w inference outputs of the remaining subjects can be found in Appendix B.

The T1w model performed better when using the manually selected mid-sagittal slices.

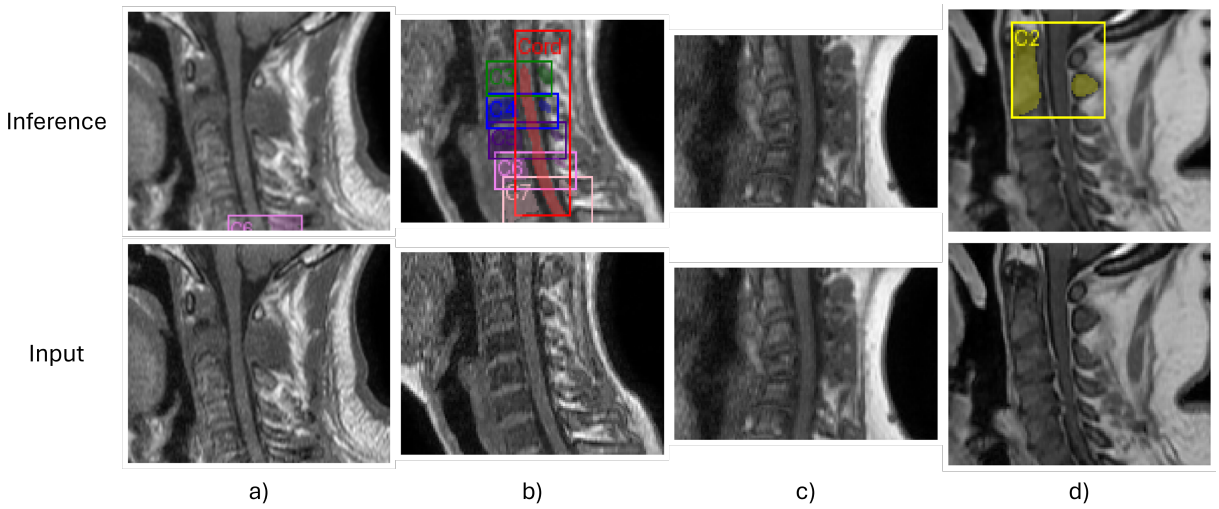


Figure 4.8: Example inference outputs of T1w MRI scans from the SpineMets clinical dataset using the mid-sagittal slice of the volume. a) Subject 1 contains C1-C4 with inference results incorrectly detecting C6. b) Subject 2 contains C2-C7 with inference results detecting C3-C7 and the spinal cord correctly. c) Subject 3 contains C2-C6 with no detections in inference. d) Subject 4 contains C1-C5 with only C2 detected correctly.

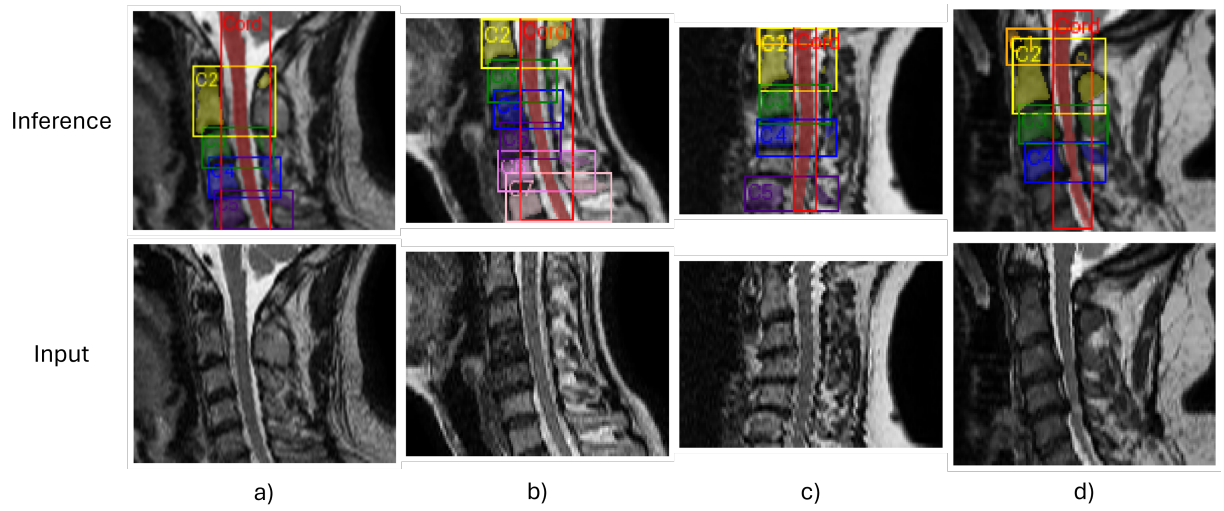


Figure 4.9: Example inference outputs of T2w MRI scans from the SpineMets clinical dataset using the mid-sagittal slice of the volume. a) Subject 1 contains C1-C5 with inference results detecting C2-C5 and the spinal cord correctly. b) Subject 2 contains C2-C7 with inference results detecting C2-C7 and the spinal cord correctly. c) Subject 3 contains C1-C6 with inference results detecting C1-C4 and the spinal cord correctly, with a mislabelled detection of C6 as C5. d) Subject 4 contains C1-C5 with C1-C4 and the spinal cord correctly detected.

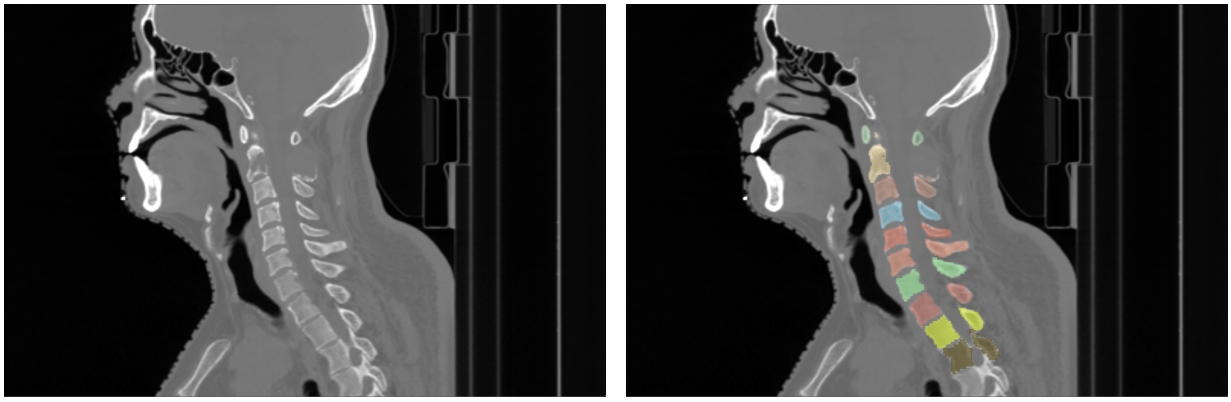


Figure 4.10: Example VertDetect inference outputs of CT scans from the SpineMets clinical dataset. a) Input CT scan on Subject 1 containing C1-T4. b) Inference results containing labelled 3D segmentations of vertebrae.

While no images had all vertebrae correctly detected with this manually selected slice, overall correct detections increased when comparing results of Subject 1 (Figure 4.8 a) and Figure 4.11 a)) and Subject 3 (Figure 4.8 c) and Figure 4.11 c)). This resulted in an increased IDR of 0.23 and IRA of 0. The best performing subject was still Subject 2 (Figure 4.11 b), where the mid-sagittal slice of the volume was found to be the optimal slice. Subject 4 (Figure 4.11 d)) had the manually selected slice performing worse, with an incorrect detection of C6 on the posterior element of C5, in comparison to the previous correct C2 detection (Figure 4.8 d)).

The T2w model performed marginally better when applied to the manually selected mid-sagittal slices, resulting in an improved IDR of 0.86 and IRA of 0.14. In Figure 4.12 a), the model was able to detect all present vertebrae, including C1 which was previously missed using the mid-sagittal slice of the volume. Figure 4.12 b) also detected all vertebrae present; however, with an erroneous detection of C1 in the air gap space. Lastly, Figure 4.12 c) and d) performed identically to the previous inference results using the mid-sagittal slice of the volume.

The performance of the model on clinical data was found to be influenced by the selection of the mid-sagittal slice used for inference as seen by the increase in IDR performance. This was expected due to the nature of the model being trained on slices mid-sagittal to the spine anatomy. However, it did not resolve the poor performance of the T1w inference, attributed to the different scanning parameters compared to the training data.

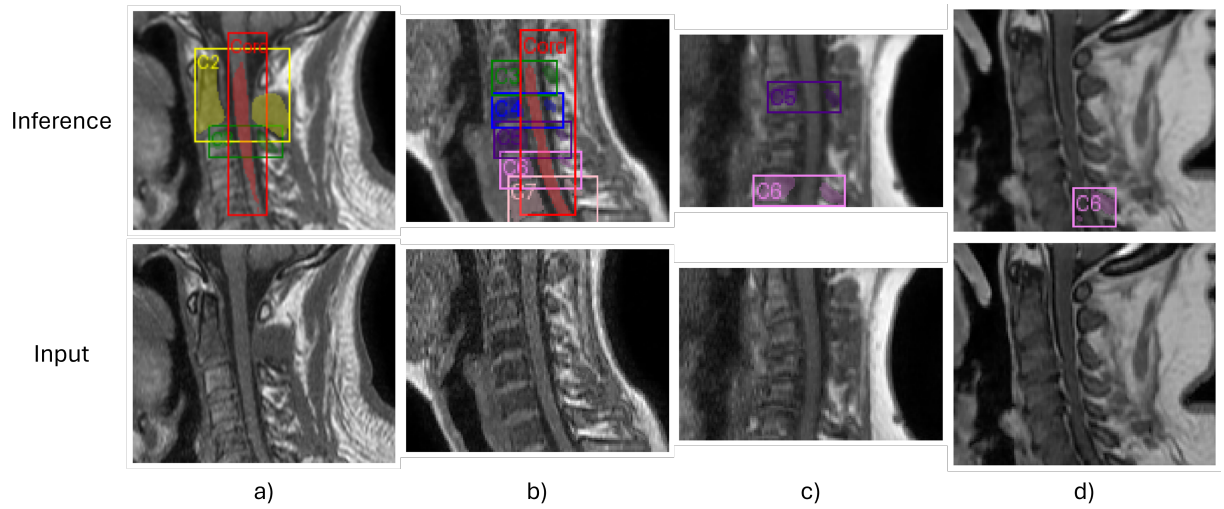


Figure 4.11: Inference outputs of T1w MRI scans from the SpineMets clinical dataset using the manually selected mid-sagittal slice. a) Subject 1 contains C1-C4 with inference results correctly detecting C2-C3 and the spinal cord. b) Subject 2 contains C2-C7 with inference results detecting C3-C7 and the spinal cord correctly. Note that the mid-sagittal slice of the volume was the ideal slice for this subject. c) Subject 3 contains C1-C6 with C6 correctly detected, and C3 incorrectly detected as C5. d) Subject 4 contains C1-C5 with an incorrect detection of C6 on the C5 posterior element.

For deformable registration, we used the inference results using the manually selected mid-sagittal due to the increased performance, in addition to only the T2w data. This allows for the evaluation of the best-case scenario of the multi-modal deformable image registration pipeline.

4.4.4 Deformable Multi-Modal Registration

Deformable multi-modal registration was then performed on the T2w SpineMets subjects. After detecting vertebrae in the CT and MRI scans, centroids were used for initial affine registration by calculating the affine transformation matrix. The CT scan was then cropped to match the 3D FoV of the MRI scan, followed by continuous optimization-based deformable registration Figure 4.13 shows the CT and MRI multi-modal affine and deformable registration process of Subject 1. Figure 4.13 a) shows the affine registered T2w volume superimposed on the CT volume, where FoV cropping to the same volumetric size and space

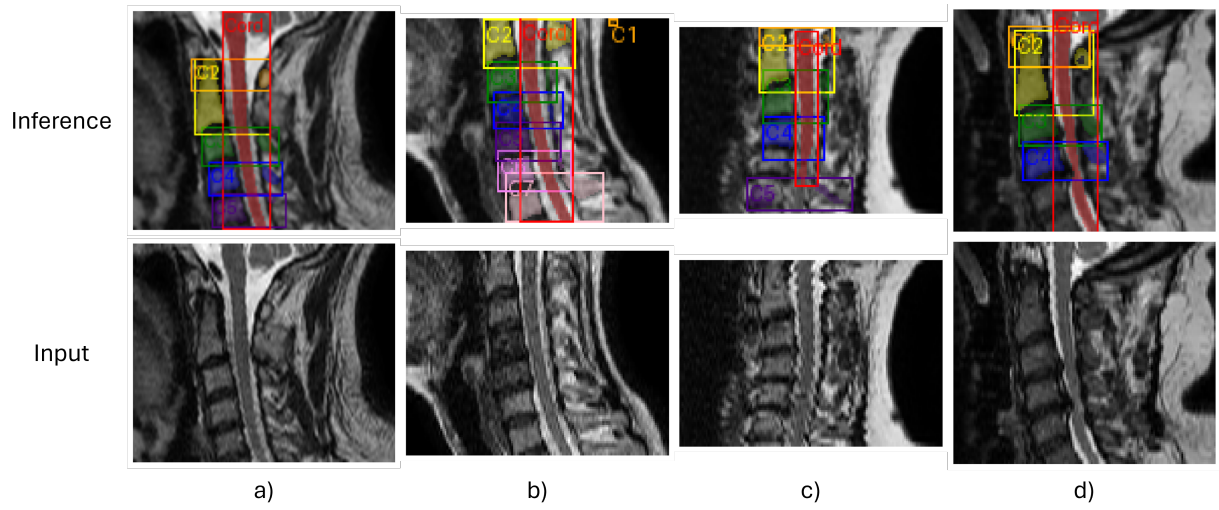


Figure 4.12: Inference outputs of T2w MRI scans from the SpineMets clinical dataset using the manually selected mid-sagittal slice. a) Subject 1 contains C1-C5 with inference results correctly detecting C1-C5 and the spinal cord. b) Subject 2 contains C2-C7 with inference results detecting C2-C7 and the spinal cord correctly, with a incorrect detection of C1 in the air gap. c) Subject 3 contains C1-C6 with inference results detecting C1-C4 and the spinal cord correctly, with a mislabelled detection of C6 as C5. Subject 4 contains C1-C5 with C1-C4 and the spinal cord correctly detected

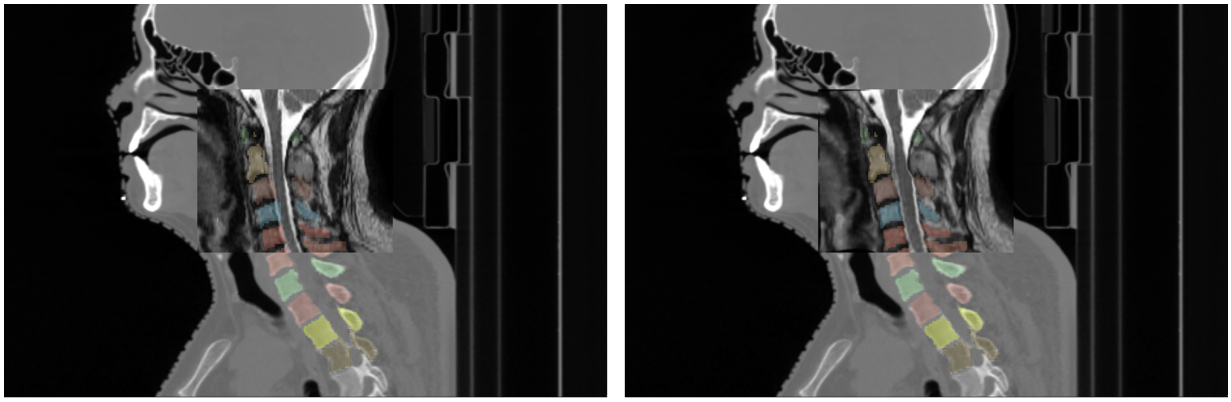


Figure 4.13: Multi-modal registration of Subject 1 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w MRI volume (C2-C5 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume. Moving T2W MRI results in vertebrae better fitting the fixed CT vertebrae segmentations.

was done as a sub-function in the pipeline. While the vertebrae are generally aligned, there are observable inconsistencies between the superimposed T2w and CT vertebrae segmentations. Figure 4.13 b) shows the successful multi-modal deformable registration applied to the moving T2w volume to the fixed CT volume. The moving T2w volume was deformed such that it better aligns with the fixed CT vertebrae segmentations. Affine and deformable registration results of remaining subjects can be found in Appendix C.

Detection Scenarios and Effects on Image Registration

Results from this preliminary analysis on clinical data showed the pipeline's ability to perform well with multi-modal deformable registration, especially for Subject 1, where all vertebrae were correctly detected. However, there were detection scenarios where not all vertebrae are correctly detected and labelled by the Mask R-CNN model. This included cases where there were: (1) undetected vertebrae, (2) mislabelled vertebrae, or (3) a combination of undetected and mislabelled vertebrae.

1. Undetected vertebra or a sequence of undetected vertebrae can be caused by the Mask R-CNN model not detecting the vertebrae or the graph filter removing detections

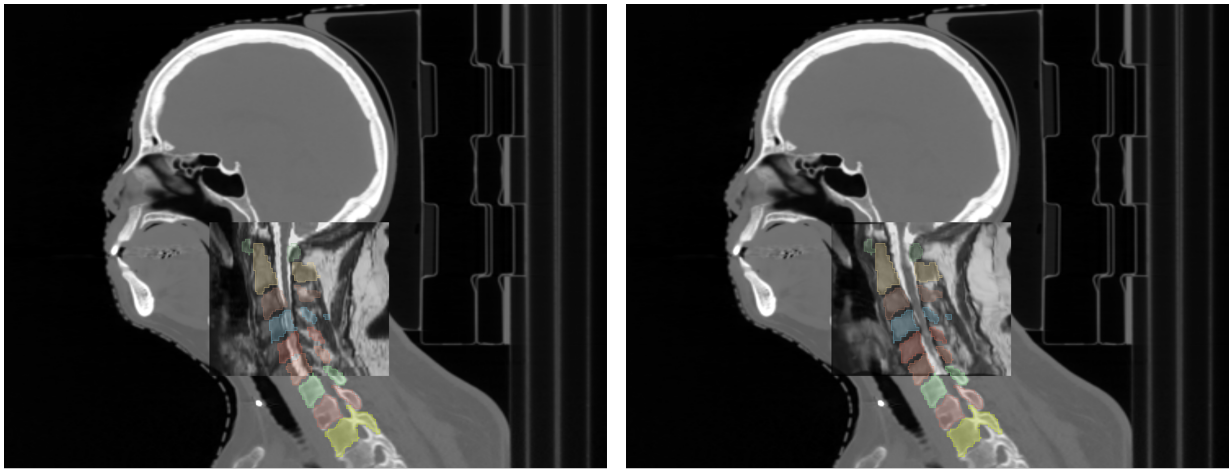


Figure 4.14: Multi-modal registration of Subject 4 with T2w volume and VertDetect vertebrae detections on CT volume. One undetected vertebrae from inference. a) Affine registration of T2w volume (C1-C4 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w Volume to CT volume. Moving T2W MRI results in vertebrae better fitting the fixed CT vertebrae segmentations even with the undetected vertebrae from inference.

due to the handcrafted rules generating the weighted directed acyclic graph (DAG). This can be observed in the detections of Subject 4 (Figure 4.12 d)). However, affine registration can still be performed to the image pair, resulting in alignment in physical space with low registration error as seen in Figure 4.14 a). Next, when FoV cropping of the CT volume was conducted, the anatomical space in the CT is also cropped such that undetected vertebrae are also kept in the cropped volume. This allows for deformable registration (Figure 4.14 b)) to be applied with minimal error in this case.

2. Mislabelled vertebrae caused by the Mask R-CNN model can result in a single vertebral level shift or the graph causing the same vertebral level shift. For example, C3 is labelled C4, C4 is labelled C5, etc. While the affine registration can be applied with minimal registration error, there will be a shift in the cropped CT vertebrae. If the shifted vertebral labels are in a region of similar looking vertebrae, it may be difficult to distinguish deformable registration errors.

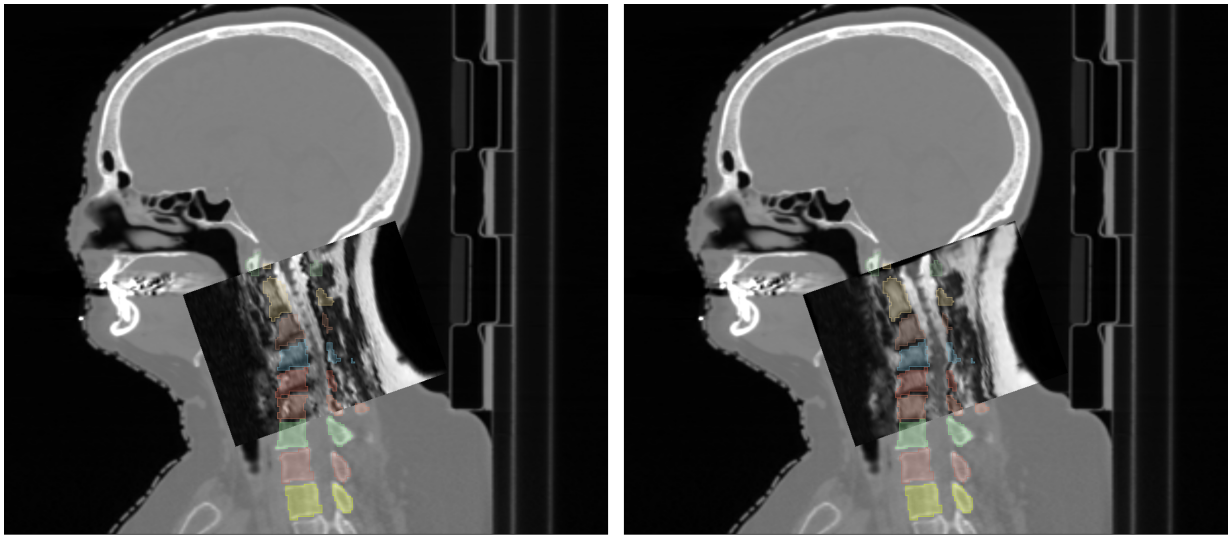


Figure 4.15: Multi-modal registration of Subject 3 with T2w volume and VertDetect vertebrae detections on CT volume. One undetected and mislabelled vertebrae from inference. a) Affine registration of T2w volume (C1-C4 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w Volume to CT volume. Moving T2W MRI results in vertebrae better fitting the fixed CT vertebrae segmentations even with the undetected and mislabelled vertebrae from inference.

3. The last scenario includes a combination of undetected and mislabelled vertebrae, where the Mask R-CNN misses the detection of a vertebrae between two detected vertebrae. The use of the graph filter then causes the inferior detection to be mislabelled, regardless if the original detection was correct. This can be observed in the vertebrae detections of Subject 3 (Figure 4.12 c)), where C5 is undetected, and C6 is labelled as C5. Unlike scenario (1), which only has a mismatch in the FoV cropping, there will be a larger affine registration error, due to the centroid of C5, mislabelled on C6, being further from the C5 centroid present in the CT image as seen in Figure 4.15. However, resulting deformable registration, as seen in Figure 4.15, still performed well.

4.5 Conclusions

This work successfully integrated the Mask R-CNN cervical vertebrae and spinal cord detection and labelling models into a multi-sequence (T1w to T2w) and multi-modal (anatomical T2w MRI to CT) deformable image registration pipeline. The Mask R-CNN model was first integrated into a multi-sequence pipeline, validated using subjects from the Spine Generic dataset unseen from training. Inference predictions from the model was utilized for FoV cropping and subsequent successful deformable registration. In addition, the Mask R-CNN model was integrated into a multi-modal deformable image registration pipeline for applications with clinical data. We successfully demonstrated the ability of the Mask R-CNN model to localize and label cervical vertebrae in the T2w clinical data for following T2w to CT deformable image registration. However, the model was unable to generalize to detect cervical vertebrae consistently in the clinical T1w images, resulting in a failure to automate T1w to CT deformable image registration. Specifically, when using a manually selected optimal slice for inference, we observe an IDR of 0.23 and 0.86 for the T1w and T2w subjects, respectively. The T1w model also failed to identify all vertebrae correctly in any image (IRA of 0), in comparison to the T2w model with an IRA of 0.14. Lastly, we successfully demonstrated multi-modal (anatomical MRI to CT) deformable registration of the cervical vertebrae using clinical data. The anatomical T2w to CT deformable registration was found to be robust to certain cervical vertebrae detection errors from the Mask R-CNN model.

Chapter 5

Thesis Conclusion

This chapter concludes the thesis and highlights the overall contributions made. The significance of research contributions, limitations of the research, and potential future research areas are discussed.

5.1 Significance of Research Contributions

In this thesis, a Mask R-CNN model was designed for detecting and labelling cervical vertebrae and the spinal cord in anatomical MRI scans. This model was trained using a custom-annotated cervical spine dataset derived from the Spine Generic dataset. Furthermore, the model was successfully incorporated into a multi-sequence and multi-modal deformable image registration pipeline. By integrating the cervical vertebrae MRI detector within this pipeline, we can identify the limiting field of view (FoV) between image pairs facilitating sufficient alignment seeding and image cropping, prior to subsequent deformable registration. In a clinical setting, this integration enables faster, more cost-effective, and accurate image registration, benefiting both pre-procedural planning and post-procedural treatment evaluation.

5.2 Limitations and Future Work

The contributions of this work provides a robust model for cervical vertebrae and spinal cord detection in MRI, in addition to integration of the model into an automated multi-

model deformable registration pipeline. However, limitations exist that require further research and development in the future.

5.2.1 Limitations and Future Work for the Mask R-CNN Cervical Vertebrae and Spinal Cord Detector

Although the Mask R-CNN model successfully demonstrated the ability to detect cervical vertebrae and the spinal cord in MRI, there were limitations that could be explored in future work.

Dataset limitations were present within the data generation process. The data generation process required the manual segmentation correction of the PAM50 warped semantic segmentation of the spine. This segmentation correction was conducted by a non-expert and posterior element segmentations would benefit from the expertise of an experienced spinal clinician or radiologist. In addition, the dataset used only contained healthy subjects imaged with a standardized protocol. This resulted in poor performance when applying the model on the T1w clinical data, where there was pathology, in addition to the significant image feature contrast differences between the clinical data and training data. Expansion of the dataset to include imaging with pathology and different imaging protocols would greatly increase model generalization and clinical application.

Inference on clinical data was restricted to subjects with images containing only the cervical spine, reducing the available dataset since many images containing cervical vertebrae also included thoracic vertebrae. Additionally, the presence of C2 in the cervical spine data limited the model's validation on smaller fields of view (FoVs). To enhance the model's applicability to clinical data, it is necessary to expand the training dataset to include other spinal regions or the entire spine.

The model was also limited to detection on 2D slices. While the model was able to generate vertebrae centroids in the volume by knowing the slice used for inference, it also limited the functionality of integrating the model into the multi-modal image registration pipeline. Specifically, the companion VertDetect model was able to predict 3D segmentations of the vertebrae, opposed to the Mask R-CNN model predicting 2D vertebrae segmentations. This limits the ability to register MRI vertebrae segmentations to the CT vertebrae segmentations. In addition, the model was sensitive to the slice used for inference, due to the anatomical mid-sagittal slice used for training. These limitations can be resolved by developing a 3D volumetric vertebrae and spinal cord detection model.

Improvements can also be explored for improving model performance. We trained the T1w and T2w model separately, employing transfer learning in the T2w model training.

Experimentation should be conducted on training a single multi-sequence cervical vertebrae and spinal cord detector. Hyper parameter tuning should also be performed to further optimized trained models.

Lastly, the graph filter faced limitations due to the use of handcrafted rules and parameters for generating the weighted directed acyclic graph (DAG) used in post-inference filtering. Although this approach worked well with the Spine Generic data, it resulted in larger errors when applied to clinical data. The rules and parameters were specifically tailored for the Spine Generic data and may need further tuning to optimize performance and applicability to clinical data.

5.2.2 Limitations and Future Work for the Multi-Modal Deformable Image Registration Pipeline

The multi-modal deformable image registration pipeline was successfully developed for automated spine image registration. Limitations of the cervical vertebrae and spinal cord detector were the primary roadblock towards applications of the pipeline to all available clinical data. However, some limitations do exist in the current pipeline as is, specifically the selection of an optimized mid-sagittal slice for model inference. We found that the slice used for inference did improve detection performance and therefore methods must be developed to automatically find the optimal slice. One method may include the segmentation of the spinal cord in 3D and using the mid-sagittal slice of the detected spinal cord volume as the input for vertebrae detection. Additionally, future work includes integrating the entire pipeline into a 3D Slicer module for streamlined usage in a clinical or research setting, and further evaluation of registration performance against clinically registered imaging from the SpineMets dataset.

References

[1] Julien Cohen-Adad, Eva Alonso-Ortiz, Mihael Abramovic, Carina Arneitz, Nicole Atcheson, Laura Barlow, Robert L. Barry, Markus Barth, Marco Battiston, Christian Büchel, Matthew Budde, Virginie Callot, Anna J. E. Combes, Benjamin De Leener, Maxime Descoteaux, Paulo Loureiro de Sousa, Marek Dostál, Julien Doyon, Adam Dvorak, Falk Eippert, Karla R. Epperson, Kevin S. Epperson, Patrick Freund, Jürgen Finsterbusch, Alexandru Foias, Michela Fratini, Issei Fukunaga, Claudia A. M. Gandini Wheeler-Kingshott, Giancarlo Germani, Guillaume Gilbert, Federico Giove, Charley Gros, Francesco Grussu, Akifumi Hagiwara, Pierre-Gilles Henry, Tomáš Horák, Masaaki Hori, James Joers, Kouhei Kamiya, Haleh Karbasforoushan, Miloš Keřkovský, Ali Khatibi, Joo-Won Kim, Nawal Kinany, Hagen H. Kitzler, Shannon Kolind, Yazhuo Kong, Petr Kudlička, Paul Kuntke, Nyoman D. Kurniawan, Slawomir Kusmia, René Labounek, Maria Marcella Laganà, Cornelia Laule, Christine S. Law, Christophe Lenglet, Tobias Leutritz, Yaou Liu, Sara Llufriu, Sean Mackey, Eloy Martinez-Heras, Loan Mattera, Igor Nestrasil, Kristin P. O'Grady, Nico Papinutto, Daniel Papp, Deborah Pareto, Todd B. Parrish, Anna Pichiecchio, Ferran Prados, Àlex Rovira, Marc J. Ruitenberg, Rebecca S. Samson, Giovanni Savini, Maryam Seif, Alan C. Seifert, Alex K. Smith, Seth A. Smith, Zachary A. Smith, Elisabeth Solana, Y. Suzuki, George Tackley, Alexandra Tinnermann, Jan Valošek, Dimitri Van De Ville, Marios C. Yiannakas, Kenneth A. Weber II, Nikolaus Weiskopf, Richard G. Wise, Patrik O. Wyss, and Junqian Xu. Open-access quantitative MRI data of the spinal cord and reproducibility across participants, sites and manufacturers. *Scientific Data*, 8(1):219, December 2021.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN, January 2018. arXiv:1703.06870 [cs].

[3] File:Illu vertebral column.svg - Wikipedia, December 2020.

[4] File:718 Vertebra-en.svg - Wikipedia, December 2015.

[5] File:Gray86.png - Wikipedia.

[6] File:Gray87.png - Wikipedia.

[7] Jeffrey G. Jarvik and Richard A. Deyo. Diagnostic Evaluation of Low Back Pain with Emphasis on Imaging. *Annals of Internal Medicine*, 137(7):586–597, October 2002. Publisher: American College of Physicians.

[8] Leonid Kalichman, David H. Kim, Ling Li, Ali Guermazi, and David J. Hunter. Computed tomography–evaluated features of spinal degeneration: prevalence, intercorrelation, and association with self-reported low back pain. *The Spine Journal*, 10(3):200–208, March 2010.

[9] Louis-Philippe Amiot and François Poulin. Computed Tomography-Based Navigation for Hip, Knee, and Spine Surgery. *Clinical Orthopaedics & Related Research*, 421:77–86, April 2004.

[10] Patrick M Colletti, Herrick J Siegel, Mary Y Woo, and Howard Y Young. THE IMPACT ON TREATMENT PLANNING OF MRI OF THE SPINE IN PATIENTS SUSPECTED OF VERTEBRAL METASTASIS: AN EFFICACY STUDY. *Computerized Medical Imaging and Graphics*, 20(3), 1996.

[11] Alberto A. Perez, Edward S. Yoon, Sravisht Iyer, Virginie Lafage, Harvinder Sandhu, Frank Schwab, Todd J. Albert, Sheeraz Qureshi, Han Jo Kim, and Yoshihiro Katsuura. Computed Tomography and Magnetic Resonance Imaging Overlay in the Spine for Surgical Planning: A Technical Report. *HSS Journal®*, 18(3):439–447, August 2022. Publisher: SAGE Publications.

[12] Gisele C. Pereira, Melanie Traughber, and Raymond F. Muzic. The Role of Imaging in Radiation Therapy Planning: Past, Present, and Future. *BioMed Research International*, 2014:e231090, April 2014. Publisher: Hindawi.

[13] Gang-Un Kim, Min Cheol Chang, Tae Uk Kim, and Gun Woo Lee. Diagnostic Modality in Spine Disease: A Review. *Asian Spine Journal*, 14(6):910–920, December 2020.

[14] Garrett K Harada, Zakariah K Siyaji, Sadaf Younis, Philip K Louie, Dino Samartzis, and Howard S An. Imaging in Spine Surgery: Current Concepts and Future Directions. *Spine Surgery and Related Research*, 4(2):99–110, November 2019.

[15] Alison Middleditch and Jean Oliver. *Functional Anatomy of the Spine*. Elsevier Health Sciences, September 2005. Google-Books-ID: y5f_UYKL28UC.

[16] Nikolai Bogduk. Functional anatomy of the spine. *FUNCTIONAL ANATOMY OF THE SPINE*, 2016.

[17] Andreas Prescher. Anatomy and pathology of the aging spine. *European Journal of Radiology*, 1998.

[18] Yogesh Kumar and Daichi Hayashi. Role of magnetic resonance imaging in acute spinal trauma: a pictorial review. *BMC Musculoskeletal Disorders*, 17(1):310, July 2016.

[19] Patrick Freund. MRI in traumatic spinal cord injury: from clinical assessment to neuroimaging biomarkers. *The Lancet Neurology*, 18, 2019.

[20] M. I. Vargas, B. M. A. Delattre, J. Boto, J. Gariani, A. Dhouib, A. Fitsiori, and J. L. Dietemann. Advanced magnetic resonance imaging (MRI) techniques of the spine and spinal cord in children and adults. *Insights into Imaging*, 9(4):549–557, August 2018. Number: 4 Publisher: SpringerOpen.

[21] Blair Allen Winegar, Matthew Derek Kay, and Mihra Taljanovic. Magnetic resonance imaging of the spine. *Polish Journal of Radiology*, 85:e550–e574, September 2020.

[22] Katarzyna Krupa and Monika Bekiesińska-Figatowska. Artifacts in Magnetic Resonance Imaging. *Polish Journal of Radiology*, 80:93–106, February 2015.

[23] Maxim Zaitsev, Julian. Maclaren, and Michael Herbst. Motion Artefacts in MRI: a Complex Problem with Many Partial Solutions. *Journal of magnetic resonance imaging : JMRI*, 42(4):887–901, October 2015.

[24] Inger Havsteen, Anders Ohlhues, Kristoffer H. Madsen, Janus Damm Nybing, Hanne Christensen, and Anders Christensen. Are Movement Artifacts in Magnetic Resonance Imaging a Real Problem?—A Narrative Review. *Frontiers in Neurology*, 8:232, May 2017.

[25] Michael A. Ibrahim, Bita Hazhirkarzar, and Arthur B. Dublin. Gadolinium Magnetic Resonance Imaging. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.

[26] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A Survey of Modern Deep Learning based Object Detection Models, April 2021.

[27] María Inmaculada García Ocaña, Karen López-Linares Román, Nerea Lete Urzelai, Miguel Ángel González Ballester, and Iván Macía Oliver. Medical Image Detection Using Deep Learning. In Yen-Wei Chen and Lakhmi C. Jain, editors, *Deep Learning in Healthcare: Paradigms and Applications*, pages 3–16. Springer International Publishing, Cham, 2020.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919.

[29] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556 [cs].

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs].

[31] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. arXiv:1905.11946 [cs, stat].

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[33] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-Time Flying Object Detection with YOLOv8, May 2024. arXiv:2305.09972.

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection, February 2018. arXiv:1708.02002.

[35] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, October 2003.

[36] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable Medical Image Registration: A Survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, July 2013.

[37] Meelis Lootus, Timor Kadir, and Andrew Zisserman. Vertebrae Detection and Labelling in Lumbar MR Images. In Jianhua Yao, Tobias Klinder, and Shuo Li, editors, *Computational Methods and Clinical Applications for Spine Imaging*, Lecture Notes in Computational Vision and Biomechanics, pages 219–230, Cham, 2014. Springer International Publishing.

[38] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 712–722, Cham, 2020. Springer International Publishing.

[39] Chenyi Zeng, Shen Zhao, Bin Chen, An Zeng, and Shuo Li. Feature-correlation-aware history-preserving-sparse-coding framework for automatic vertebra recognition. *Computers in Biology and Medicine*, 160:106977, June 2023.

[40] Shen Zhao, Xiangsheng Li, Jiayi He, Bin Chen, and Shuo Li. Sequence based local–global information fusion framework for vertebrae detection under pathological and FOV variation challenges. *Computerized Medical Imaging and Graphics*, 108:102244, September 2023.

[41] Daniel Forsberg, Erik Sjöblom, and Jeffrey L. Sunshine. Detection and Labeling of Vertebrae in MR Images Using Deep Learning with Clinical Annotations as Training Data. *Journal of Digital Imaging*, 30(4):406–412, August 2017.

[42] Ben Glocker, J. Feulner, Antonio Criminisi, D. R. Haynor, and E. Konukoglu. Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7512, pages 590–598. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[43] Ben Glocker, Darko Zikic, Ender Konukoglu, David R. Haynor, and Antonio Criminisi. Vertebrae Localization in Pathological Spine CT via Dense Classification from Sparse Annotations. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 262–270, Berlin, Heidelberg, 2013. Springer.

[44] Dong Yang, Tao Xiong, Daguang Xu, Qiangui Huang, David Liu, S. Kevin Zhou, Zhoubing Xu, JinHyeong Park, Mingqing Chen, Trac D. Tran, Sang Peter Chin, Dimitris Metaxas, and Dorin Comaniciu. Automatic Vertebra Labeling in Large-Scale 3D CT using Deep Image-to-Image Network with Message Passing and Sparsity Regularization, May 2017. arXiv:1705.05998 [cs].

[45] Geoff Klein, Michael Hardisty, Cari Whyne, and Anne L. Martel. VertDetect: Fully End-to-End 3D Vertebral Instance Segmentation Model, November 2023. arXiv:2311.09958 [cs, eess].

[46] Mohammed Benjelloun, Saïd Mahmoudi, and Fabian Lecron. A Framework of Vertebra Segmentation Using the Active Shape Model-Based Approach. *International Journal of Biomedical Imaging*, 2011(1):621905, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2011/621905.

[47] Isaac Castro-Mateos, Jose M. Pozo, Marco Pereañez, Karim Lekadir, Aron Lazary, and Alejandro F. Frangi. Statistical Interspace Models (SIMs): Application to Robust 3D Spine Segmentation. *IEEE Transactions on Medical Imaging*, 34(8):1663–1675, August 2015. Conference Name: IEEE Transactions on Medical Imaging.

[48] M. Hardisty, L. Gordon, P. Agarwal, T. Skrinskas, and C. Whyne. Quantitative characterization of metastatic disease in the spine. Part I. Semiautomated segmentation using atlas-based deformable registration and the level set method. *Medical Physics*, 34(8):3127–3134, August 2007.

[49] Amin Suzani, Alexander Seitel, Yuan Liu, Sidney Fels, Robert N. Rohling, and Purang Abolmaesumi. Fast Automatic Vertebrae Detection and Localization in Pathological CT Scans - A Deep Learning Approach. In Joachim Hornegger, Nassir Navab, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 678–686. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.

[50] Shen Zhao, Xi Wu, Bo Chen, and Shuo Li. Automatic Vertebrae Recognition from Arbitrary Spine MRI Images by a Hierarchical Self-calibration Detection Framework. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 316–325, Cham, 2019. Springer International Publishing.

[51] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. SpineNetV2: Automated Detection, Labelling and Radiological Grading Of Clinical MR Scans, May 2022. arXiv:2205.01683 [cs, eess].

[52] Renjie Wang, Jennifer Hui Yi Voon, Da Ma, Setareh Dabiri, Karteek Popuri, and Mirza Faisal Beg. Vertebra Segmentation for Clinical CT Images Using Mask R-CNN. In Tomaz Jarm, Aleksandra Cvetkoska, Samo Mahnič-Kalamiza, and Damijan Miklavcic, editors, *8th European Medical and Biological Engineering Conference*, IFMBE Proceedings, pages 1156–1165, Cham, 2021. Springer International Publishing.

[53] Yiqiang Zhan, Dewan Maneesh, Martin Harder, and Xiang Sean Zhou. Robust MR Spine Detection Using Hierarchical Learning and Local Articulated Model. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Nicholas Ayache, Hervé Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, volume 7510, pages 141–148. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. Series Title: Lecture Notes in Computer Science.

[54] Heyou Chang, Shen Zhao, Hao Zheng, Yang Chen, and Shuo Li. Multi-vertebrae Segmentation from Arbitrary Spine MR Images Under Global View. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 702–711, Cham, 2020. Springer International Publishing.

[55] Julien Cohen-Adad, Eva Alonso-Ortiz, Mihael Abramovic, Carina Arneitz, Nicole Atcheson, Laura Barlow, Robert L. Barry, Markus Barth, Marco Battiston, Christian Büchel, Matthew Budde, Virginie Callot, Anna J. E. Combes, Benjamin De Leener, Maxime Descoteaux, Paulo Loureiro De Sousa, Marek Dostál, Julien Doyon, Adam Dvorak, Falk Eippert, Karla R. Epperson, Kevin S. Epperson, Patrick Freund, Jürgen Finsterbusch, Alexandru Foias, Michela Fratini, Issei Fukunaga, Claudia A. M. Gandini Wheeler-Kingshott, Giancarlo Germani, Guillaume Gilbert, Federico Giove, Charley Gros, Francesco Grussu, Akifumi Hagiwara, Pierre-Gilles Henry, Tomáš Horák, Masaaki Hori, James Joers, Kouhei Kamiya, Haleh Karbasforoushan, Miloš Keřkovský, Ali Khatibi, Joo-Won Kim, Nawal Kinany, Hagen Kitzler, Shannon Kolind, Yazhuo Kong, Petr Kudlička, Paul Kuntke, Nyoman D. Kurniawan, Slawomir Kusmia, René Labounek, Maria Marcella Laganà, Cornelia Laule, Christine S. Law, Christophe Lenglet, Tobias Leutritz, Yaou Liu, Sara Llufriu, Sean Mackey, Eloy Martinez-Heras, Loan Mattera, Igor Nestrasil, Kristin P. O'Grady, Nico Papinutto, Daniel Papp, Deborah Pareto, Todd B. Parrish, Anna Pichiecchio, Ferran Prados, Àlex Rovira, Marc J. Ruitenberg, Rebecca S. Samson, Giovanni Savini, Maryam Seif, Alan C. Seifert, Alex K. Smith, Seth A. Smith, Zachary A. Smith, Elisabeth Solana, Yuichi Suzuki, George Tackley, Alexandra Tinnermann, Jan Valošek, Dimitri Van De Ville, Marios C. Yiannakas, Kenneth A. Weber, Nikolaus Weiskopf, Richard G. Wise, Patrik O. Wyss, and Junqian Xu. Generic acquisition protocol for quantitative MRI of the spinal cord. *Nature Protocols*, 16(10):4611–4632, October 2021.

[56] Benjamin De Leener, Simon Lévy, Sara M. Dupont, Vladimir S. Fonov, Nikola Stikov, D. Louis Collins, Virginie Callot, and Julien Cohen-Adad. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. *NeuroImage*, 145:24–43, January 2017.

[57] Charley Gros, Benjamin De Leener, Atef Badji, Josefina Maranzano, Dominique Eden, Sara M. Dupont, Jason Talbott, Ren Zhuoquiong, Yaou Liu, Tobias Granberg, Russell Ouellette, Yasuhiko Tachibana, Masaaki Hori, Kouhei Kamiya, Lydia Chougar, Leszek Stawiarz, Jan Hillert, Elise Bannier, Anne Kerbrat, Gilles Edan, Pierre Labauge, Virginie Callot, Jean Pelletier, Bertrand Audoin, Henitsoa Rasoanandrianina, Jean-Christophe Brisset, Paola Valsasina, Maria A. Rocca, Massimo Filippi, Rohit Bakshi, Shahamat Tauhid, Ferran Prados, Marios Yiannakas, Hugh Kearney, Olga Ciccarelli, Seth Smith, Constantina Andrada Treaba, Caterina Mainero, Jennifer Lefeuvre, Daniel S. Reich, Govind Nair, Vincent Auclair, Donald G. McLaren, Allan R. Martin, Michael G. Fehlings, Shahabeddin Vahdat, Ali Khatibi, Julien Doyon, Timothy Shepherd, Erik Charlson, Sridar Narayanan, and Julien Cohen-Adad. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *NeuroImage*, 184:901–915, January 2019.

[58] Benjamin De Leener, Vladimir S. Fonov, D. Louis Collins, Virginie Callot, Nikola Stikov, and Julien Cohen-Adad. PAM50: Unbiased multimodal template of the brainstem and spinal cord aligned with the ICBM152 space. *NeuroImage*, 165:170–179, January 2018.

[59] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V. Miller, Steve Pieper, and Ron Kikinis. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, 30(9):1323–1341, November 2012.

[60] Teodora Vujovic. Continuous Optimization with Piecewise-Decaying Learning Rate Scheduling for Medical Image Registration. *Journal of Computational Vision and Imaging Systems*, 8(1):78–80, 2022. Number: 1.

APPENDICES

Appendix A

Mask R-CNN Detection Model Training and Validation Curves

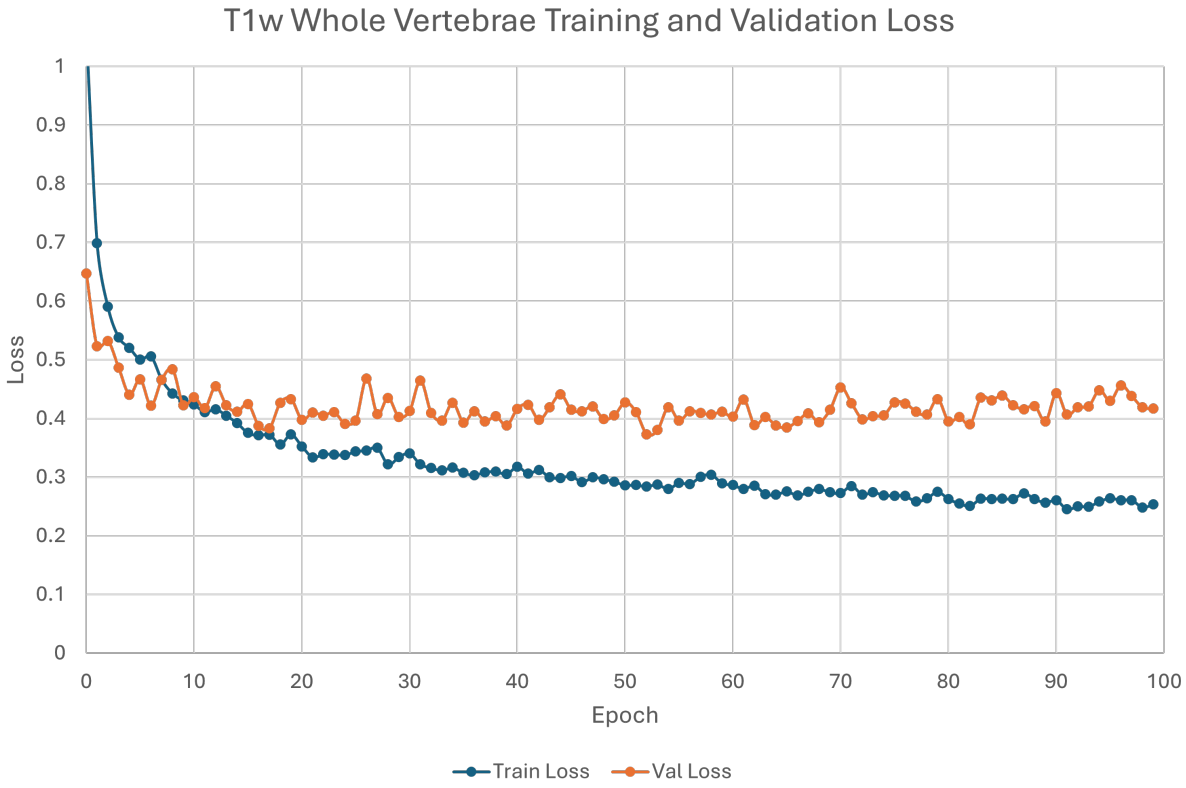


Figure A.1: T1w whole vertebrae model training and validation loss curves.

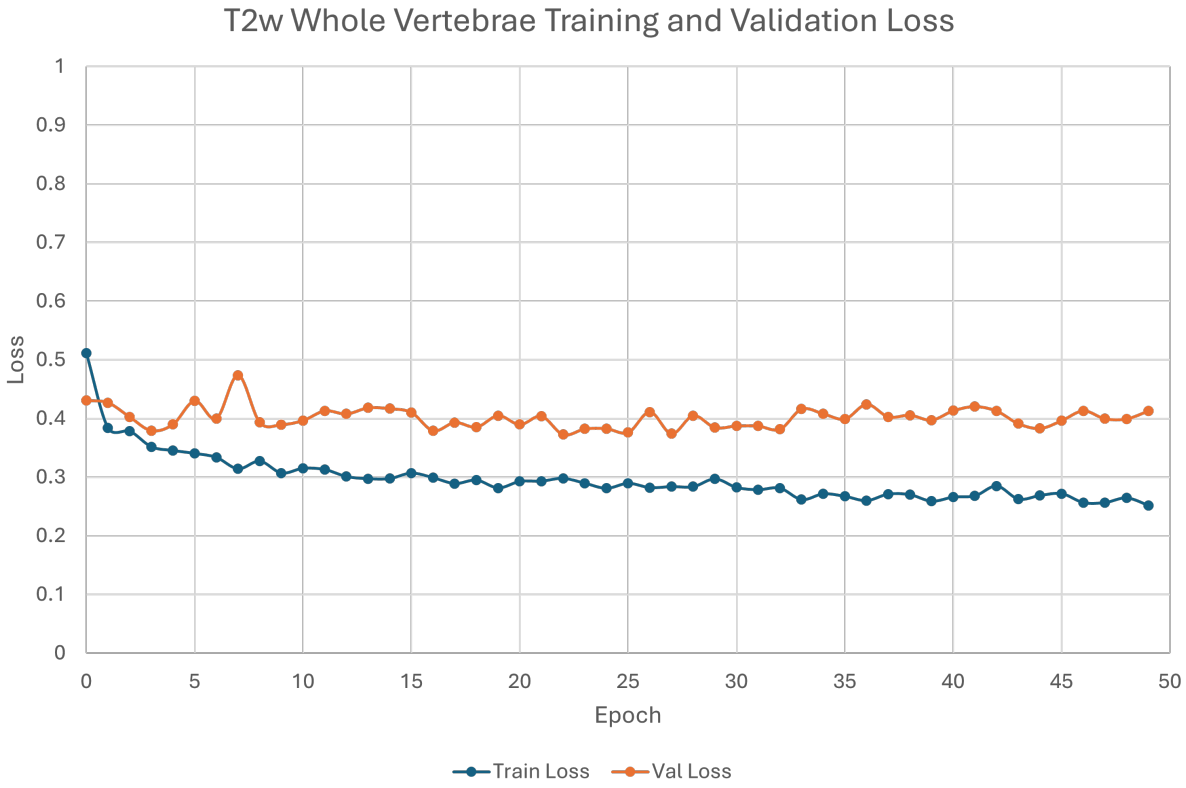


Figure A.2: T2w whole vertebrae model training and validation loss curves.

Appendix B

SpineMets MRI Inference Results

B.1 Mid-Sagittal Slice of MRI Volume Input

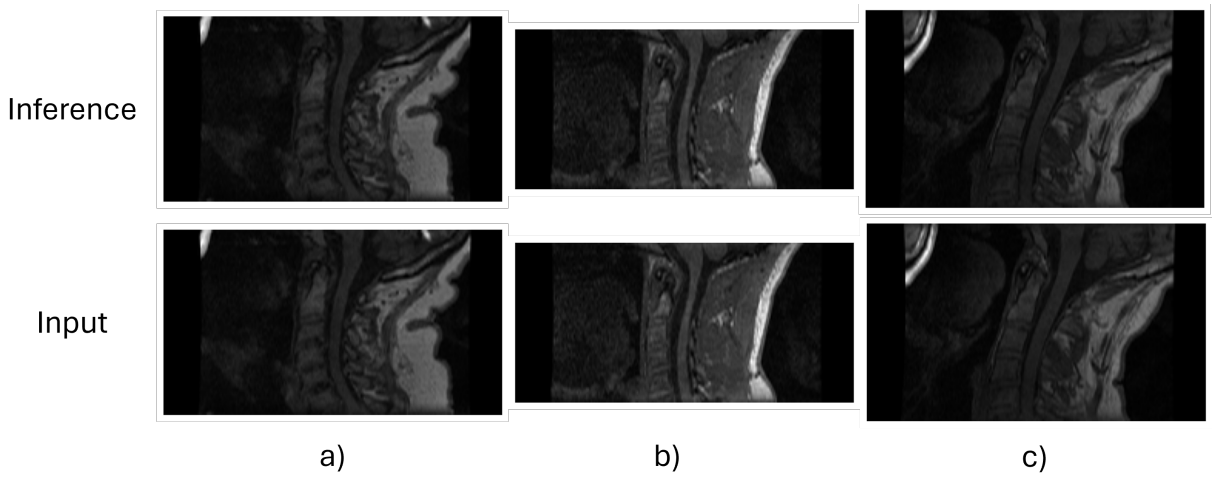


Figure B.1: Inference outputs of T1w MRI scans SpineMets clinical dataset using the mid-sagittal slice of the volume of remaining subjects. a) Subject 5 contains C1-C6 with no detections. b) Subject 6 contains C1-C5 with no detections. c) Subject 7 contains C1-C5 with no detections.

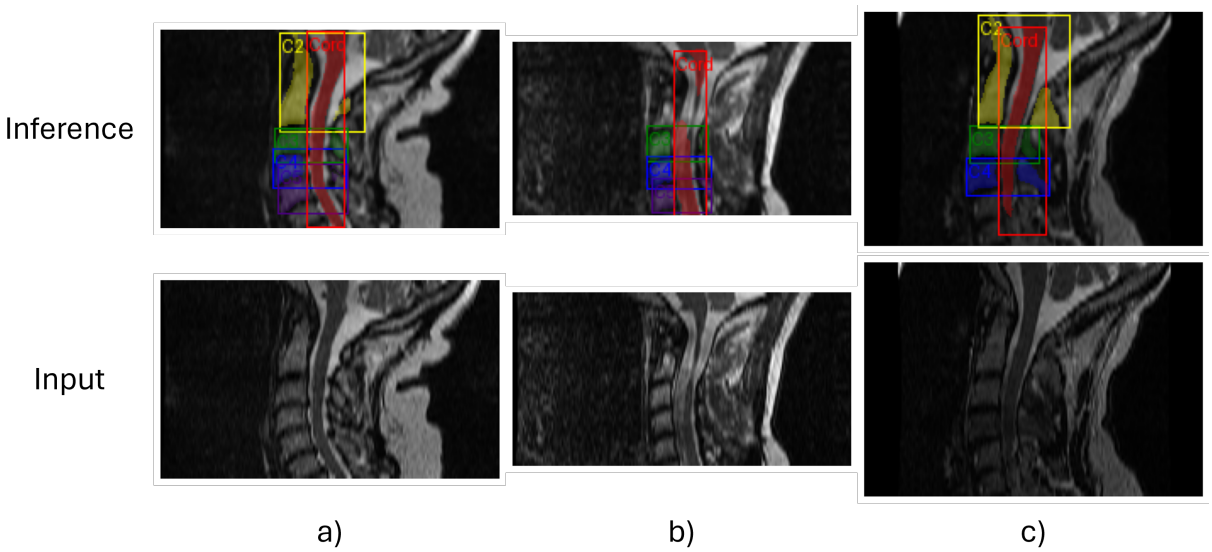


Figure B.2: Inference outputs of T2w MRI scans SpineMets clinical dataset using the mid-sagittal slice of the volume of remaining subjects. a) Subject 5 contains C2-C6 with inference results correctly detecting C2-C5 and the spinal cord, with C6 undetected. b) Subject 6 contains C2-C5 with detections of C3-C5 and the spinal cord. c) Subject 7 contains C1-C5 with C3-C4 and the spinal cord correctly detected. C2 is correctly detected, but the segmentation and bounding box extends past the vertebrae into soft tissue in the superior direction.

B.2 Manually Selected Mid-Sagittal Slice of MRI Volume Input

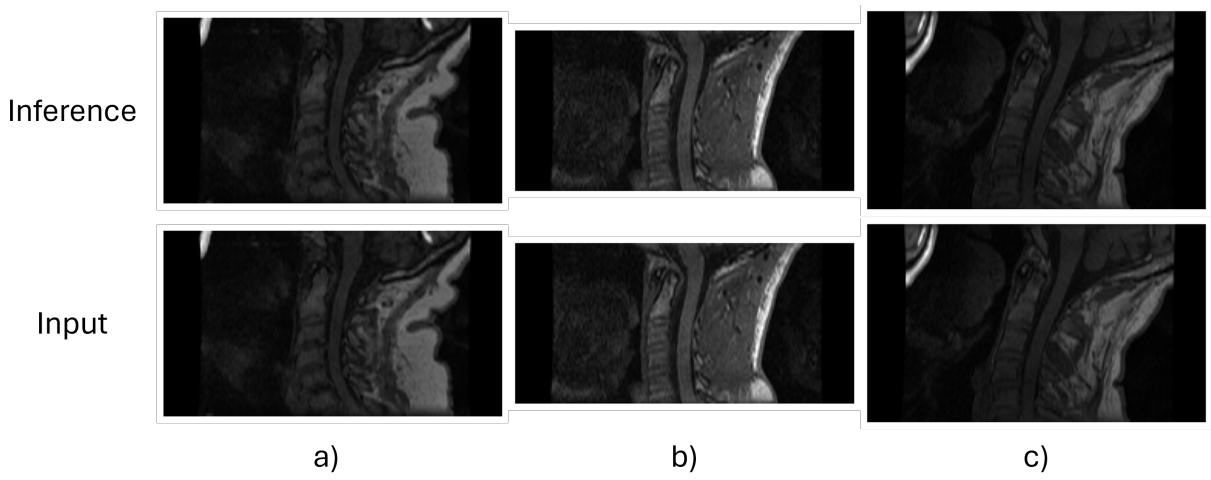


Figure B.3: of T1w MRI scans from the SpineMets clinical dataset using the manually selected mid-sagittal slice. a) Subject 5 contains C1-C6 with no detections. b) Subject 6 contains C1-C5 with no detections. d) Subject 7 contains C1-C5 with no detections.

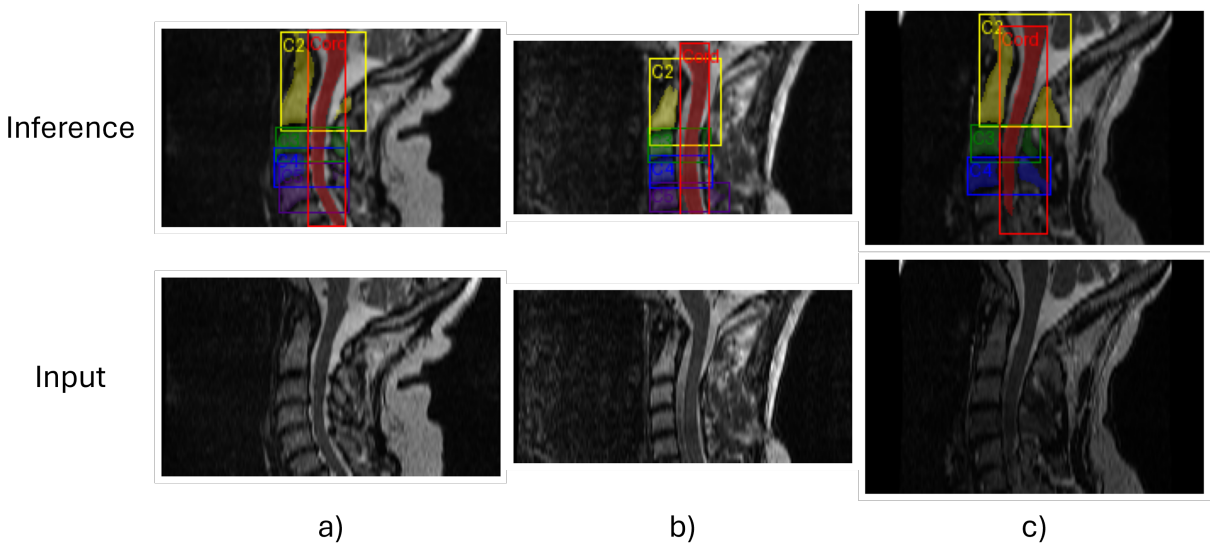


Figure B.4: Inference outputs of T2w MRI scans SpineMets clinical dataset using the manually selected mid-sagittal slice of remaining subjects. a) Subject 5 contains C2-C6 with inference results correctly detecting C2-C5 and the spinal cord, with C6 undetected. b) Subject 6 contains C2-C5 with detections of C2-C5 and the spinal cord. c) Subject 7 contains C1-C5 with C3-C4 and the spinal cord correctly detected. C2 is correctly detected, but the segmentation and bounding box extends past the vertebrae into soft tissue in the superior direction.

Appendix C

SpineMets T2w Affine and Deformable Registration Results

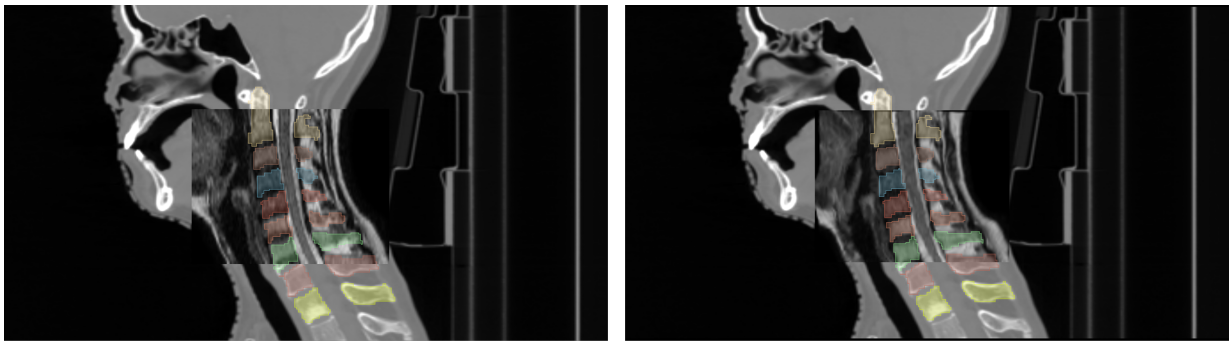


Figure C.1: Multi-modal registration of Subject 2 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C7 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume.

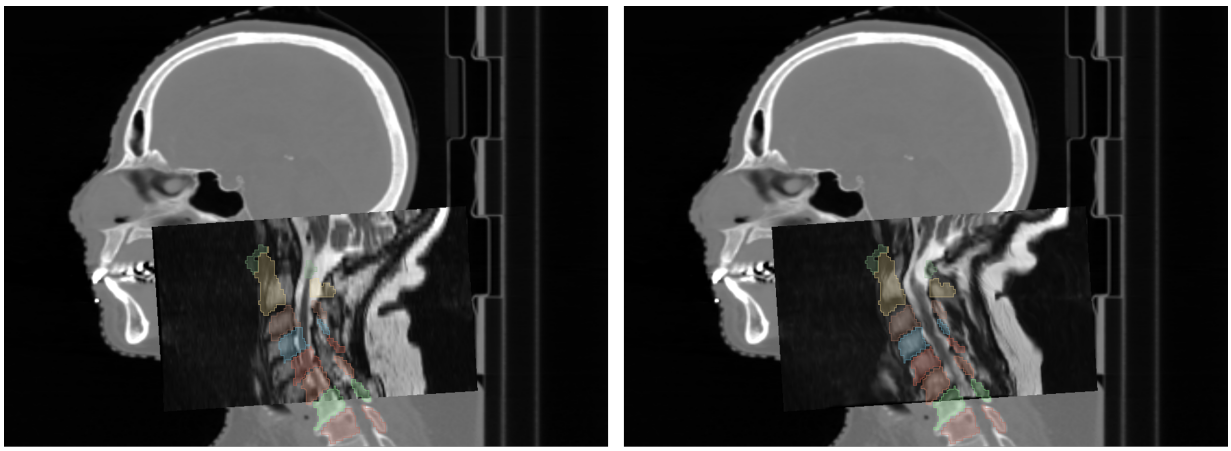


Figure C.2: Multi-modal registration of Subject 5 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C5 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume.

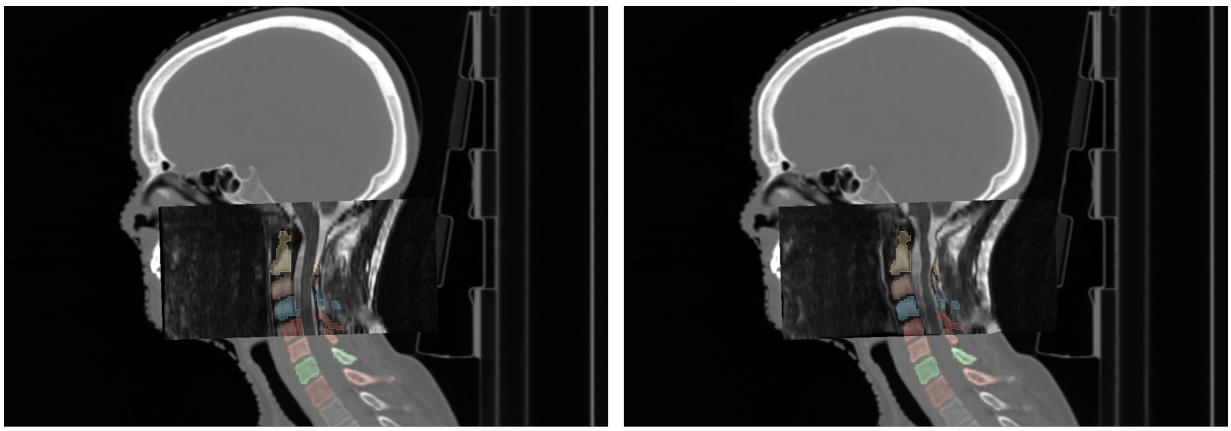


Figure C.3: Multi-modal registration of Subject 6 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C5 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume.

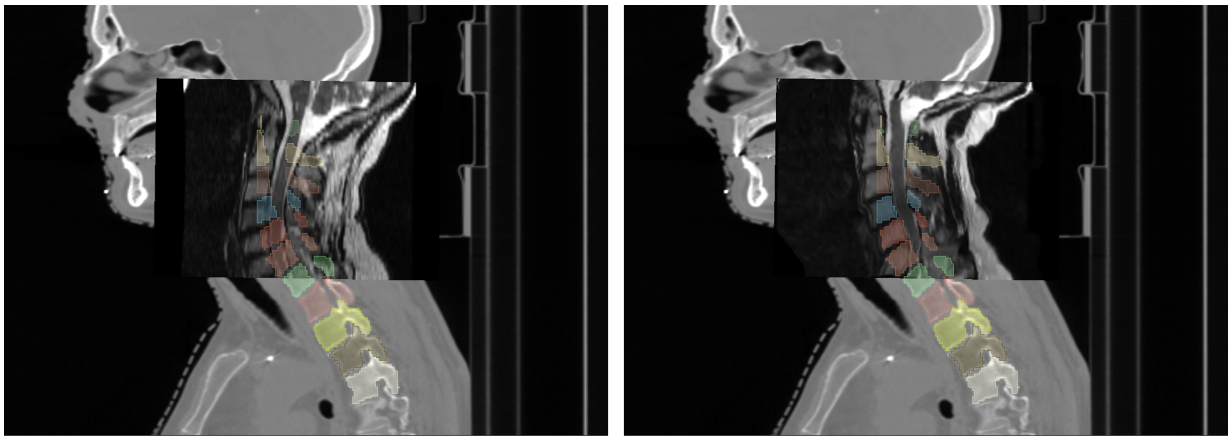


Figure C.4: Multi-modal registration of Subject 7 with T2w volume and VertDetect vertebrae detections on CT volume. a) Affine registration of T2w volume (C2-C4 detected) to CT volume and vertebrae segmentations. b) Deformable registration of T2w volume to CT volume.