Unsupervised Three-dimensional Segmentation of Scoliotic Spines from MR Volumes with Uncertainty Estimation

by

Muni Venkata Naga Karthik ENAMUNDRAM

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS IN ELECTRICAL ENGINEERING
M.A.Sc.

MONTREAL, MAY 5TH, 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Catherine Laporte, Thesis Supervisor
Department of Electrical Engineering, École de techologie supérieure

Mrs. Farida Cheriet, Co-supervisor
Department of Computer Engineering and Software Engineering, Polytechnique Montréal

Mr. Stéphane Coulombe, President of the Board of Examiners
Department of Software and IT Engineering, École de techologie supérieure

Mrs. Rita Noumeir, Independent Examiner
Department of Electrical Engineering, École de techologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MAY 31ST, 2021

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

Needless to say, the circumstances during my master's studies have been extraordinary - from lockdowns to online classes to working from home. The list just goes on and on.

First and foremost, I am deeply thankful to my supervisor Professor Catherine Laporte for her immense support and guidance during this difficult time. I owe most of this work to her faith in me, particularly in the directions I had wished to pursue during the course of my thesis project. It has been a true privilege to be supervised by her and I would be remiss to not acknowledge the positive influence her training would have on my growth as a researcher. I thank her for her time and effort spent on my project and can only admire her patience in correcting numerous drafts of my work.

I am also thankful to my co-supervisor Professor Farida Cheriet for her presence and involvement during our meetings about my project. Her insightful suggestions have helped improve the quality of this work. Participating in her lab's weekly meetings has been a fun (and important) getaway for me. I would also like to thank her research associate, Philippe Debanné, for all his efforts in retrieving the data. Important parts of the data used in this project have come from Farida's lab, and it would not have been possible if not for Philippe's help in this regard.

I am also grateful to all the other students in Cathy's research group, Jawad Dahmani, Arnaud Brignol, Sen Li, Sahba Changizi, and David Olivier, for their attention and interest during my presentations at our lab meetings. Explaining my work and introducing new concepts to them during these meetings has immensely helped in keeping my morale high.

I would like to thank Natural Sciences and Engineering Research Council (NSERC) and Fonds de recherche du Québec - Nature et technologies (FRQNT) for their financial support. I also thank Compute Canada for generously offering their massive computational resources for this work. I would like to thank ÉTS for their various forms of financial support during my studies and also the accomplished technicians at the IT department for promptly resolving all the technical issues.

Lastly, the ongoing pandemic has taught me the importance of human connection. I am deeply thankful to my parents for their never-ending love and moral support.

Segmentation tri-dimensionnelle non supervisée de la colonne vertébrale scoliotique de volumes d'IRM avec estimation de l'incertitude

Muni Venkata Naga Karthik ENAMUNDRAM

RÉSUMÉ

La segmentation des vertèbres à partir d'images de résonance magnétique (IRM) est une tâche difficile. En raison de la nature de cette modalité, qui met l'accent sur les tissus mous du corps, les algorithmes de seuillage courants sont inefficaces pour détecter les os dans les IRM. D'autre part, il est relativement plus facile de segmenter les os à partir d'images de tomographie axiale (CT) en raison du contraste élevé entre les os et les tissus mous.

Cette thèse propose une nouvelle méthode de segmentation simple basée sur le seuillage de la colonne vertébrale scoliotique qui génère son modèle 3D en effectuant une synthèse inter-modalités entre les domaines IRM et CT. Cependant, cela suppose implicitement la disponibilité de données IRM-CT appariées, ce qui est rare, surtout dans le cas de patients scoliotiques. Par conséquent, la méthode proposée est totalement non supervisée et entièrement tridimensionnelle (3D). Un modèle CycleGAN 3D est entraîné pour une traduction non appariée de volume à volume dans les domaines de l'IRM et de la CT. Ensuite, l'algorithme de seuillage d'Otsu est appliqué aux volumes CT synthétisés pour une segmentation facile des os vertébraux. La segmentation résultante est utilisée pour reconstruire un modèle 3D de la colonne vertébrale. La nature non supervisée du problème ainsi que le manque de données CT de base rendent difficile l'analyse objective de la performance du CycleGAN 3D. Par conséquent, une adaptation bayésienne de CycleGAN est proposée pour capturer l'information d'incertitude sous la forme d'incertitudes aléatoires et épistémiques pendant la traduction. Cette amélioration rend le modèle autosuffisant en fournissant une mesure de la confiance dans les prédictions du modèle tout en le rendant plus interprétable pour la tâche de segmentation d'Otsu.

Des expériences de type étude d'ablation ont été menées pour déterminer l'importance de l'estimation de l'incertitude et l'amélioration de la qualité des volumes traduits est également démontrée. La méthode proposée est validée quantitativement sur 46 vertèbres scoliotiques de 5 patients présentant différents degrés de courbure de la colonne vertébrale en calculant la distance moyenne point-surface entre les points de repère de chaque vertèbre obtenus à partir de radiographies préopératoires et la surface de la vertèbre segmentée. L'étude aboutit à une erreur moyenne de 3.17 ± 1.04 mm. Sur la base des résultats qualitatifs et quantitatifs, il est conclu que le cadre proposé est capable d'obtenir de bonnes segmentations et de bons modèles 3D de la colonne vertébrale scoliotique ainsi que des informations cruciales sur l'incertitude, le tout après un apprentissage à partir de données non appariées de manière non supervisée.

Mots-clés: Segmentation des vertèbres, scoliose, synthèse multimodale, CycleGAN 3D, incertitude

Unsupervised Three-dimensional Segmentation of Scoliotic Spines from MR Volumes with Uncertainty Estimation

Muni Venkata Naga Karthik ENAMUNDRAM

ABSTRACT

Vertebral bone segmentation from magnetic resonance (MR) images is a challenging task. Due to the inherent nature of the modality to emphasize soft tissues of the body, common thresholding algorithms are ineffective in detecting bones in MR images. On the other hand, it is relatively easier to segment bones from computed tomography (CT) images because of the high contrast between bones and the surrounding regions.

This thesis proposes a novel method for a simple thresholding-based segmentation of the scoliotic spine that generates its 3D model by performing a cross-modality synthesis between MR and CT domains. However, this implicitly assumes the availability of paired MR-CT data, which is rare, especially in the case of scoliotic patients. Therefore, the proposed method is completely unsupervised and fully three-dimensional (3D). A 3D CycleGAN model is trained for an unpaired volume-to-volume translation across MR and CT domains. Then, the Otsu thresholding algorithm is applied to the synthesized CT volumes for easy segmentation of the vertebral bones. The resulting segmentation is used to reconstruct a 3D model of the spine. The unsupervised nature of the problem along with the lack of ground truth CT data make it difficult to objectively analyze the performance of the 3D CycleGAN. Therefore, a Bayesian adaptation of CycleGAN is proposed for capturing the uncertainty information in the form of aleatoric and epistemic uncertainties during translation. This enhancement makes the model self-sufficient by providing a measure of the confidence in the model's predictions while simultaneously making it more interpretable for the Otsu segmentation task.

Ablation-study type experiments were run to determine the importance of uncertainty estimation and the improvement in the quality of translated volumes is also shown. The proposed method is quantitatively validated on 46 scoliotic vertebrae in 5 patients with varying degrees of the spinal curvature by computing the point-to-surface mean distance between the landmark points for each vertebra obtained from pre-operative X-rays and the surface of the segmented vertebrae. The study results in a mean distance error of 3.17 ± 1.04 mm. Based on qualitative and quantitative results, it is concluded that the proposed framework is able to obtain good segmentations and 3D models of the scoliotic spines along with the crucial uncertainty information, all after training from unpaired data in an unsupervised manner.

Keywords: Vertebrae segmentation, Scoliosis, Cross-modality synthesis, 3D CycleGAN, Uncertainty

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
3D	Three-dimensional
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
CMS	Cross Modality Synthesis
NCC	Normalized Cross Correlation
ICP	Iterative Closest Point
KL	Kullback-Liebler
GP	Gaussian Process
NLL	Negative Log-Likelihood
MR	Magnetic Resonance
GC	Gradient Consistency

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

T	Tesla
mm	millimeters

INTRODUCTION

Motivation

Scoliosis is a complex three-dimensional (3D) deformity of the human trunk. Along with a lateral deviation in the spine and an axial rotation of the vertebrae, a deformation in the rib cage is also observed. For decades, X-ray images have been the gold standard in the initial diagnosis of scoliosis with systematic radiographic imaging performed over the course of the treatment to regularly monitor the growth of the spinal curvature. Severely afflicted patients are also required to undergo surgical treatment to correct the spinal curvature. The posterior spinal fusion and instrumentation is the surgical procedure that applies to most scoliosis patients, wherein, pedicle screws and hooks are inserted to anchor the implanted rods. A spinal fusion then helps in welding the bone grafts and the vertebrae into a solid mass. The pedicle screw insertion during surgery carries a high risk of spinal cord injury with possible nerve root damage. It has also been shown that intra-operative image guidance using computed tomography (CT) images can improve the surgical accuracy by lowering breach rates when compared to freehand surgical methods (Chan et al., 2019). However, the radiation dose absorbed from the CT scans poses a greater risk and therefore, precludes its common usage. On the other hand, magnetic resonance imaging (MRI) provides a reliable, non-invasive alternative to CT scans by producing cross-sectional, 3D images of the body *without* irradiating the patient. MR images capture the soft tissue regions such as the intervertebral disks in high detail and also provide good visualizations of the spinal cord and nerve roots.

In order to reduce the severity of the spinal curvature, different types of surgeries are performed. In particular, scoliosis patients having stiff curvature undergo intervertebral disk resection so that instrumentations can be attached to the rods inserted and prevent other complications. However, the removal of intervertebral disks based on 2D intraoperative images is difficult for surgeons due to the lack of depth perception and the loss of geometric structural information. In

such cases, computer-assisted diagnosis can help in improving surgical accuracies by providing additional contextual information in the form of distance between the surgical tools and the highly-critical spinal cord. Thanks to the suitable characteristics of MR imaging mentioned above, MR volumes can be used to segment vertebral column and 3D models of the spine can be reconstructed. Thus, these 3D preoperative spine models could then be registered to intraoperative image data for surgical assistance (Chevrefils et al., 2009).

Another issue with the treatment of scoliosis is that, while the corrective measures such as the application of orthotic braces or a spine surgery result in the optimal shape of the spine, they do not necessarily correspond to the optimal external shape of the trunk (which concerns the appearance of the patient). A study by Harmouche et al. (2012) shows that one can simulate the effect of the spine surgery on the external shape of the trunk by combining the soft tissue information from MRIs and the spine information from X-rays to obtain a complete 3D model of the patient trunk. This is done by propagating the effect of the spinal correction to the trunk surface by extracting the soft-tissue information from MR data. The advantage is that the 3D model accounts for the postural differences between the MR and X-ray data (as the former is acquired in prone position, while latter is acquired in standing position) and also includes the soft tissue.

Recent studies (D'Andrea et al. (2015), De Silva et al. (2017)) have shown that preoperative MR can be co-registered to intraoperative CT and radiographic images respectively, for real-time navigation during spine surgeries, thereby improving the overall rate of surgical success. The former used a dataset of adults ranging from ages 18-75. However, there is an intensity mismatch and tissue non-correspondence between the MR and CT images. While the task of vertebral level localization and segmentation is easier in pre- and intra-operative CT co-registration due to superior bone detail, a direct adaption of such solutions for preoperative MR and intraoperative CT co-registration is challenging and hence their outcomes remain largely uncertain. As a

consequence, it is difficult for automatic segmentation algorithms to accurately detect the bone structures in MR images. Therefore, this suggests that for an effective utilization of preoperative MR images during surgical treatments, the accuracy of the vertebral segmentation and localization, despite the challenges, is crucial. One possible method is to take advantage of the superior bones intensities in CT images while leveraging the recent advancements in field of deep learning (DL) to *artificially synthesize* the pixel intensities of CT images from their MR counterparts.

Successful breakthroughs in the field of artificial intelligence (AI) have paved the way for an ever-increasing application of DL techniques that span across several domains including, but not limited to, computer vision, natural language processing and medical imaging. In vision problems, the superiority of AI in imitating human-level performance in pattern recognition tasks is largely attributed to the increasing availability of large and labelled datasets for training (ImageNet (Deng et al., 2009), CIFAR100 (Krizhevsky, 2012), etc.) and the decreasing cost and ease of access to huge computational resources. While it is natural to expect a similar trend in the medical domain, it is not difficult to realize that the breakthroughs in the application of DL techniques for medical image analysis have not been on par with its vision counterparts. Some of the important reasons are as follows:

- **Lack of training data**. While the advancements in technology have definitely transformed healthcare in general, the problem of insufficient data acquisition still exists. In some cases, it is simply because some pathologies cannot be imaged as easily as others, and much less so compared to the acquisition of natural images in vision problems. Moreover, the time and expertise required for labeling the acquired images is expensive.
- **Underwhelming generalization of successful vision models**. In most cases, a successful vision model is adapted for its application to medical images. Due to the drastic difference in the content of natural and medical images, a dip in the performance is seen. One way

in which this can be avoided is by designing DL models specific to the pathology under consideration by working in close collaboration with physicians.

- **Ethical concerns of medical data**. Privacy is an important factor concerning medical data, especially. The patients may or may not choose to publicly release their data under certain conditions, which is the main reason why one cannot find many large publicly available medical datasets.

A clever method for tackling the lack of training data, and the one that is discussed in-depth in this thesis, is the concept of *cross-modality synthesis*. In simple terms, this concept allows for the artificial synthesis of medical images across two domains - MRI to CT, CT to Ultrasound, etc. to name a few. This has been largely possible due to the popularity of a specific type of DL model called the generative adversarial network (GAN) (Goodfellow et al., 2014). GANs have gained huge attention in recent years because of their capability of synthesizing data with unprecedented levels of realism.

While the prospect of artificially synthesizing medical images to tackle one of the longstanding problems in medical imaging is exciting, it is crucial to also understand the consequences that it entails. DL models have infamously been known as *black-box* models due to their inability to provide interpretable and explainable results. This is particularly crucial in the case of medical imaging because the decisions taken by the clinicians as a result of the DL models' performance have the ability to directly affects the patients' lives. Hence, it is important to deeply understand the shortcomings of these models, instead of taking their typically overconfident point estimate outputs (such as classification accuracies) as gospel. In this regard, Bayesian deep learning offers mathematically principled methods for analyzing the performance of DL models in depth. This thesis discusses a specific extension called *Bayesian uncertainty estimation*, in the context of unsupervised medical image synthesis. By extracting uncertainty estimates, we can understand

where the model is under-confident (or, overconfident) in its predictions, thereby providing the clinicians with interpretable results that can help in improving overall patient care.

Problem Statement

The objective is to obtain a 3D segmentation of the spine and reconstruct its 3D model, given only the preoperative MR volumes of scoliotic patients without the corresponding CT volumes.

Therefore, we tackle the following problems in this thesis:

- Given how easily the vertebral bones can be segmented from CT images, **how can we utilize the advancements in cross-modality synthesis to *translate* the bone intensities across the MR and CT domains?**
- Assuming a successful cross-domain MR $\rightarrow$ CT translation, **can the vertebral bones then be easily segmented using an off-the-shelf segmentation algorithm?**
- **Most importantly**, as the corresponding CT data for scoliotic patients are unavailable, **can this be achieved in an *unsupervised* way?**
- Given the unsupervised learning problem, to what extent can we trust the model's predictions? Is there any meaningful way to **quantify the uncertainty in an unsupervised model?**

Proposed Solution

The contributions in this thesis are as follows:

1. A deep-learning-based image synthesis method is proposed for vertebral column segmentation and 3D model reconstruction given only a patient's preoperative scoliotic MR volume. There are two important points that should be emphasized:
    - While the idea of cross-modality synthesis was introduced from an image-to-image translation perspective for simplicity, in practice, a *volume-to-volume* translation is performed - going from MR volumes to CT volumes of the scoliotic spines,

- A major challenge is the unavailability of paired MR and CT volumes for a given patient. This is an important issue because the lack of ground truth data puts the onus on the model itself to become self-sufficient in terms of obtaining correct translations. Therefore, this constraint forced us to use an unsupervised volume synthesis method.

2. Owing to the unsupervised nature of the problem, a novel method for quantifying the uncertainty depending on the model and the data on which is trained upon, is also proposed. This manifests itself in the form of additional interpretable information to the user in the regions where the model's confidence in its prediction is low.

The method consists of three main stages, which augmented by a computationally-efficient intermediate step for modelling aleatoric (data-dependent) uncertainty and epistemic (model-dependent) uncertainty within the proposed framework. The novelty of this method lies in the fact that this is the first approach combining both aleatoric and epistemic uncertainties in an *unsupervised* learning setting. These stages will be explained in further detail in chapter 3. They are briefly defined as follows:

1. **Synthesis with Uncertainty** - We propose an augmented version of 3D CycleGAN, a powerful variant of GANs, for translating the bone intensities from CT volumes to the MR volumes with uncertainty estimation. In particular, modeling uncertainty is shown to be useful in two ways:
   - The model-dependent uncertainty provides additional interpretable information to the user for the segmentation task in terms of showing the regions translated with low confidence (high uncertainty).
   - The data-dependent uncertainty helps the model improve its performance by learning to differentiate between the regions surrounding the spine during the course of training.
2. **Segmentation** - The Otsu segmentation algorithm is applied to the synthesized CT volume to perform the segmentation. We note that Otsu's method is a commonly known thresholding-based segmentation algorithm.

3. **Reconstruction** - The segmented volume is used for reconstructing a 3D model of the scoliotic spine.

The rest of the thesis is structured as follows:

- Chapter 1 educates the reader with sufficient background knowledge thereby laying a foundation for the major contents of this thesis. A detailed overview of scoliosis, an introduction to CT and MR imaging along with their advantages and disadvantages with regards to vertebral bone segmentation, a brief review of convolutional neural networks (CNNs), and a thorough mathematical review of GANs are discussed in this chapter.
- Chapter 2 covers the literature surrounding our problem. A wide array of previously published results mainly involving image processing, graph-based and classical ML methods, is reviewed. The types of image synthesis methods using GANs is presented along with the recent literature in Bayesian DL and uncertainty estimation. This chapter also highlights the lack of literature involving the segmentation of scoliotic spines and the novelty in the proposed solution is shown.
- Chapter 3 is the core of this thesis, giving a thorough overview of the proposed methodology for vertebral bone segmentation from MR volumes and uncertainty estimation. The theory behind the quantitative validation approach used in this study is also discussed.
- Chapter 4 presents the qualitative and quantitative results along with a discussion on how modeling uncertainty helps in vertebral bone segmentation.
- Finally, a brief review of the major contributions is given and a few directions for future work are discussed.

The preliminary work leading to this thesis was published in the Proceedings of the SPIE Medical Imaging Conference titled, *Three-dimensional Segmentation of the Scoliotic Spine from MRI using Unsupervised, Volume-based MR-CT Synthesis* (Naga Karthik, Laporte & Cheriet, 2021).

CHAPTER 1

BACKGROUND

The aim of this chapter is to lay a solid foundation for understanding the contents of the rest of the thesis by providing sufficient background knowledge to the reader on scoliosis (sections 1.1 and 1.2), CT and MR imaging (section 1.3), CNNs (section 1.4) and GANs (section 1.5).

1.1 The Vertebral Column

In order to understand why scoliosis presents an important problem, it is helpful to understand the structure of the human vertebral column. The vertebral column, also known as the spinal column or backbone, is comprised of a segmented series of bones extending from the neck to the tail bone. The *segments* are made up of individual bones, called the vertebrae and are separated by intervertebral discs. One of its function is to protect the spinal cord, which is housed inside the spinal canal. As seen in figure 1.1, there are 33 vertebrae in total that are divided into 5 regions, naturally forming the curvature of the spine. Starting from the top, there are 7 cervical vertebrae (C1-C7), 12 thoracic vertebrae (T1-T12), 5 lumbar vertebrae (L1-L5), which form the upper part of the spinal column each separated by an intervertebral disk. The remaining 9 vertebra are divided between the sacrum containing 5 fused vertebrae (S1-S5) and the coccyx containing 4 small, fused vertebrae.

The shape of a typical vertebra is complex (see figure 1.2). It consists of the vertebral body and the vertebral arch. They combine together to form the vertebral foramen, which houses the spinal cord. Out of the 7 processes that are supported by the vertebral arch, the spinous and transverse processes are of particular importance, as shall be seen in the later chapters. They are situated posterior to the vertebral body, on the center (coming out) and on the left and right each, respectively.

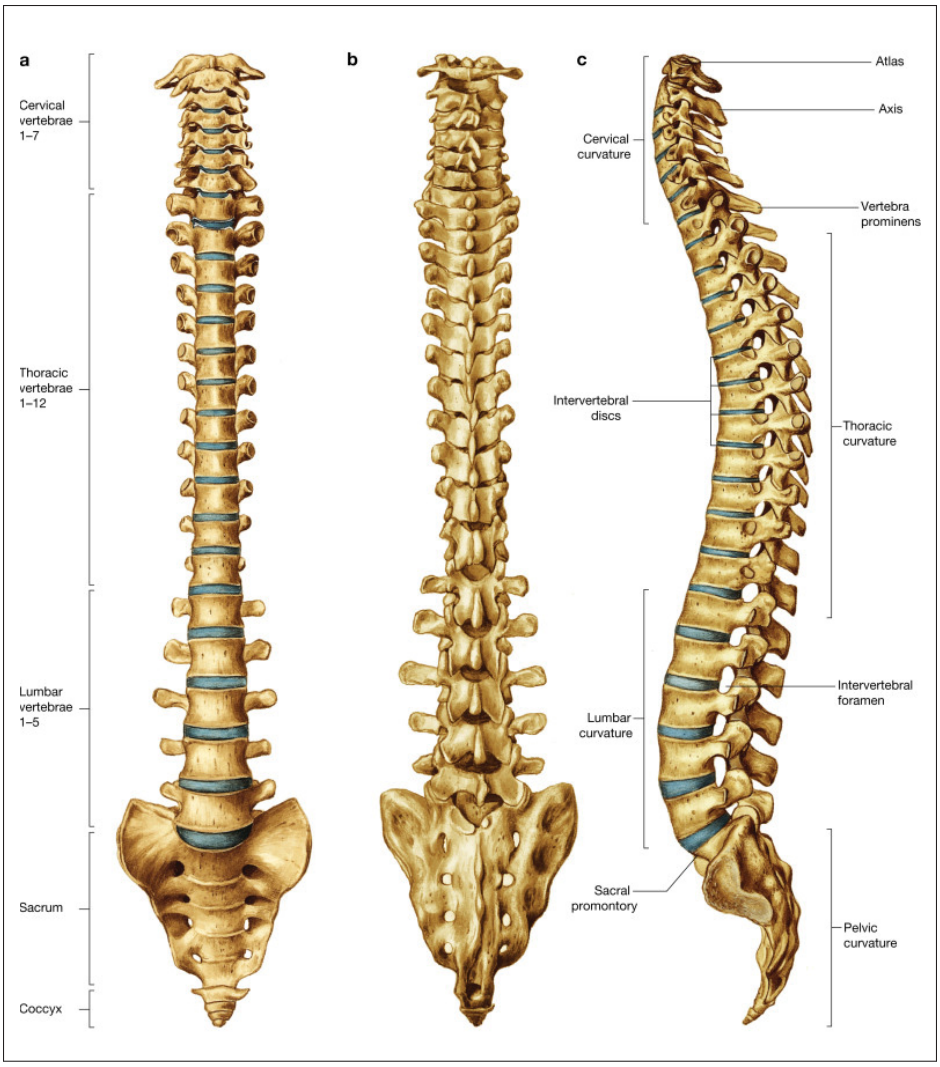


Figure 1.1 The anterior, posterior and lateral views of the human vertebral column
Taken from Mahadevan (2018)

1.2 Scoliosis

Scoliosis is a complex 3D deformity of the spine in the sagittal and coronal planes along with vertebral rotation in the axial plane. It is a medical condition in which the lateral curvature of the spine is greater than 10°. The most common method to measure the spine curvature is called the Cobb angle (figure 1.3b). It is measured from the postero-anterior X-ray images of the patient's spine acquired every 4 - 6 months. Therefore, the patients' repeated exposure to

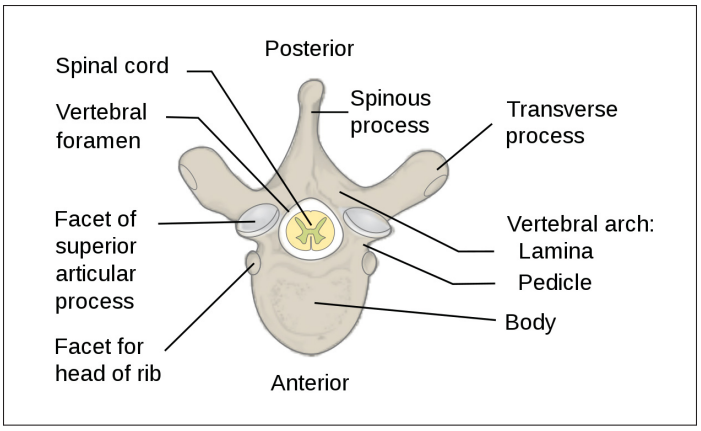
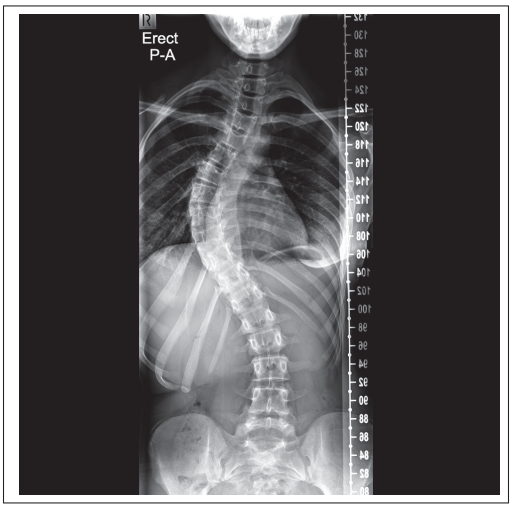


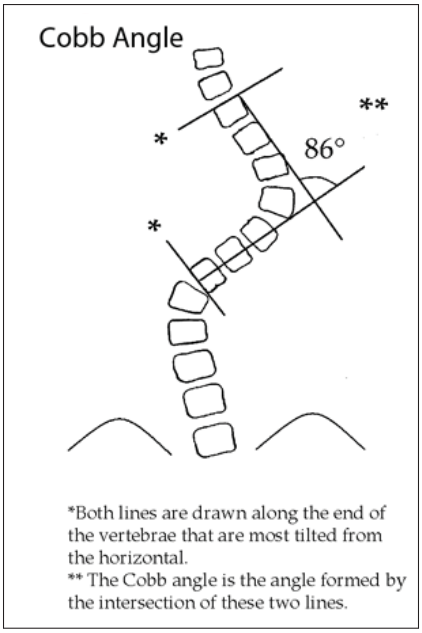
Figure 1.2 Anatomy of a vertebra. Taken from https://en.wikipedia.org/wiki/Vertebra (Consulted from Wikipedia in April 2021)

harmful radiations during each checkup also poses a health concern (Wong et al., 2019). Figure 1.3a shows the radiograph of a scoliosis patient whose Cobb angle was measured to be 65°. As seen in figure 1.3b, the apex vertebra is initially fixed and the most tilted vertebrae above and below the apex are chosen. The Cobb angle is then measured by computing the angle between the intersecting lines drawn perpendicular to the top and the bottom vertebrae.

The treatment for scoliosis generally varies according to the severity of the spinal curvature as characterized by the Cobb angle. Curves between $10° - 25°$ are considered to be mild and do not require treatment apart from the regular X-ray checkups. For curves between $25° - 40°$, the use of an orthotic brace is recommended to limit the growth of the curvature. For curvatures $> 40°$, aggressive action is required, with surgery being the recommended choice of action. There are several scenarios where MRI is required. For instance, when the patients is under 10 years old, MRI scanning is done to confirm the presence of scoliosis. MRI is also used when there is a sign of rapid curve progression or when the patient experiences atypical characteristics of scoliosis such as numbness, loss of reflexes, etc. Irrespective of the scenarios mentioned above, an MRI evaluation is always performed before a spinal fusion surgery. This, therefore, explains the ubiquity of MRI data for scoliotic patients. On the other hand, CT scans, although more detailed than X-ray images, are rarely used in scoliosis due to the high radiation risk that they



a) Radiograph of a scoliotic patient
Taken from Radiopaedia (2021)



b) Cobb angle measurement
Taken from
Daubs & Watkins-Castillo (2021)

Figure 1.3 A typical radiograph along with the technique for measuring the Cobb's angle

entail. Special cases only include the situations where MRI scans cannot be acquired due to the patient having a pacemaker or a metal implant.

Since the probability of having an abnormal bone complexity is extremely low, this naturally presents a scenario where a given scoliosis patient only undergoes MRI evaluations during the course of his/her treatment. As a result, it is quite difficult to acquire a paired MR-CT dataset for scoliotic patients, which suggests developing novel methods considering unpaired data instances only. Consistent with the previous statement, we could not find many sources in the literature that talk about vertebrae segmentation in scoliotic patients using DL methods. We believe that this is mainly because due to a conspicuous lack of large and labeled datasets that are generally required for training DL methods. We discuss this further in chapter 2.

1.3 Computed Tomography and Magnetic Resonance Imaging

1.3.1 Computed Tomography (CT) Imaging

A CT scan is a medical imaging technique that combines the slices of multiple X-ray scans taken from different angles revolving around the patient to produce a cross-sectional (tomographic) view of the body. The acquired slices are collected and digitally stacked to form a 3D image, which allows for an easy identification of the bone structures as well as other abnormalities such as tumours. In every CT scanning machine, there exists a doughnut-shaped circular structure called the gantry. It contains an X-ray source that continuously emits X-rays while rotating around the patient. The rays that leave the patients are captured by the detector and sent to the computer. After all X-ray projections in one revolution are acquired, the computer then uses complex mathematical techniques to generate a 2D cross-sectional slice, in a process known as tomographic reconstruction. This process is repeated until the desired number of slices are acquired to form a reliable 3D image of the patient's anatomy.

By the nature of their construction, CT scans are better in creating highly detailed anatomical images than planar projection X-ray images. However, this comes at the cost of exposing patients to more ionizing radiation. For comparison, the average radiation dose resulting from an X-ray scan is about $0.01 - 0.15$ mGy, whereas, it is about $10 - 20$mGy for a typical CT scan (with a few specialized CT scans such as head or cardiac CT, irradiating upto 80 mGy of radiation dosage per scan) (Hall & Brenner, 2008). It is important to note that the risk of developing radiogenic cancer from CT scans is extremely low; however, the radiation doses are accumulated and therefore, acquiring CT scans regularly becomes a major risk.

In order to avoid excessive exposure, CT acquisitions are only done when it is known that a successful scan will result in more benefits that outweigh the risks. Despite these limitations, CT scans are quite useful for certain parts of the body. For instance, they are used to detect tumours and lesions in the abdomen, a cardiac CT scan reveals the possibility of heart diseases or abnormalities, etc. In particular, CT scans are most useful in detecting bone fractures, tumours

and eroded bone-joints. Figure 1.4 (right) show the sagittal view of a spine CT scan. It can be seen that bones are prominently visible in such images, thereby facilitating their segmentation through the application of simple, intensity-based segmentation algorithms.

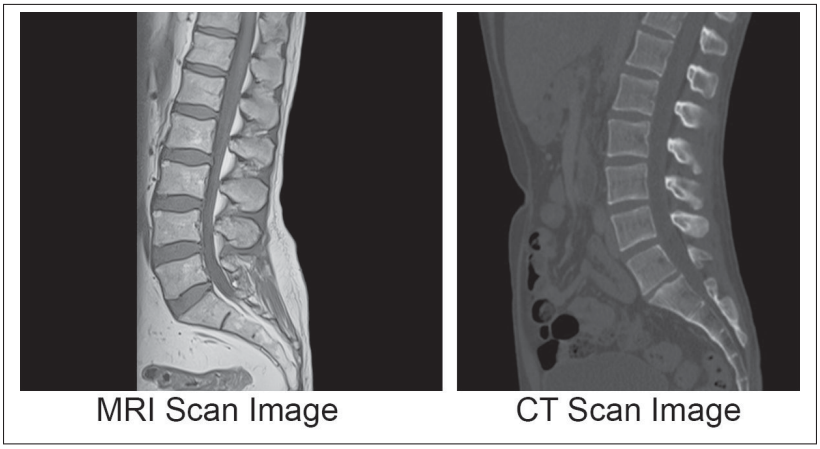


Figure 1.4 Examples of the sagittal views of the spine in MRI and CT images
Taken from Carter (2021)

1.3.2 Magnetic Resonance Imaging (MRI)

MRI is a *non-invasive* medical imaging technique that uses strong magnetic fields and radio-frequency waves to produce 3D anatomical images. These are well suited for imaging soft-tissue (i.e. non-bony) regions of the body and unlike CT, they do not use X-rays, thereby carrying no risk of ionizing radiation. Soft tissues such as the spinal cord, nerves, muscles, ligaments are clearly seen in MRI images and this is also the recommended imaging method when frequent scans need to be performed. The strength of the magnetic fields is measured in "teslas" (T). Most of the MRI scanning systems in hospitals and medical research clinics operate with a magnetic field setting of 1.5T or 3T.

The physics of MRI are an interesting phenomenon. MRI primarily exploits the fact that the hydrogen nuclei in our body's tissues can be excited to emit a signal, which can then be processed in terms of an image by capturing the density of the protons in those particular regions. Initially, a strong magnetic field is created around the patient so that all the hydrogen atoms are aligned in

the direction of the magnetic field. The protons are then excited by passing a radiofrequency current through the body. These protons, now stimulated, spin out of equilibrium and are subject to additional strain against the magnetic field's pull. When the radiofrequency current is turned off, the protons realign with the direction of the magnetic field, during the process of which, they emit energy captured by the coils in the MRI machine. The faster the realignment, the brighter is the resulting image. Tissues can be differentiated easily depending on the time taken for the protons to realign.

The strength of the magnetic fields employed in MRI systems can also be a cause for concern. Patients with metal or iron implants cannot undergo an MRI scan as those objects pose a high risk to the patients under a strong magnetic field. A characteristic repetitive noise is also heard during the scan, which can go as high as 120 decibels, requiring some ear protection. However, its main advantage is that it is non-invasive and can produce detailed images of the body. Figure 1.4 (left) shows a typical MRI scan of the spine. Note that bones are of the same intensity as their surroundings, unlike the CT scan where the bones are prominent in the foreground. This prevents simple segmentation algorithms to readily detect the bone structures, thereby suggesting the requirement for an alternative method to detect bones from MR images.

1.4 Convolutional Neural Networks

Before diving into the fundamentals of GANs, it is important to understand the building blocks that constitute any GAN. These are called convolutional neural networks (CNNs) (Le Cun et al., 1989). As the name suggests, CNNs are simply the class of neural networks that use *convolution* (instead of matrix multiplication) as the fundamental mathematical operation[1]. There exist a variety of applications where CNNs have been used successfully, ranging from 1D time-series data to multi-dimensional data. Let us briefly look at the CNN architecture in more detail.

[1] The reader is referred to Appendix I for a short review of the mathematical formula behind convolutions.

1.4.1 Architecture Overview

The primary constituent of a CNN is the convolutional layer which introduces a kernel (typically, a 3×3 or 5×5 matrix) that is slid over the entire input image. The overlapping elements of the image and the kernel are element-wise multiplied and the summation of one such multiplication represents one pixel of the new image. The summation of all such multiplications results in a new image (also known as the *feature map*). It is important to note that the kernel is initialized with random values during each instantiation of a convolutional layer and its elements are learned during training.

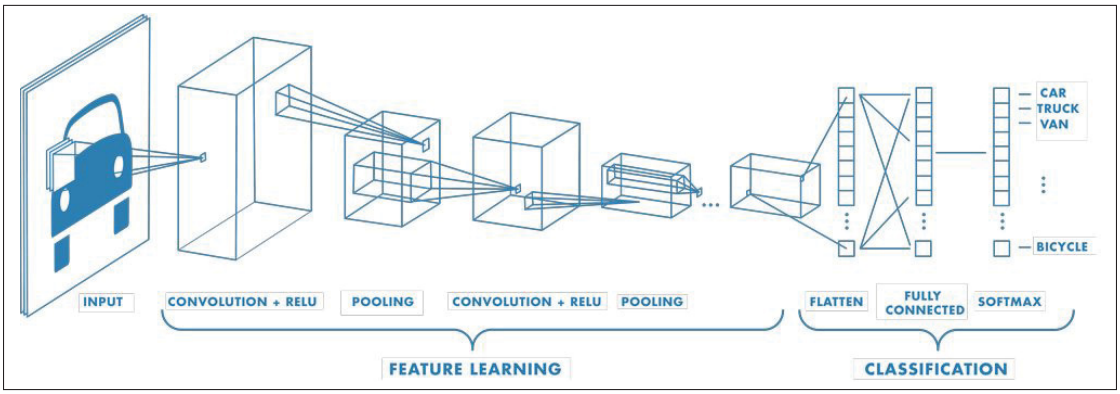


Figure 1.5 A typical CNN architecture with all of its layers.
Taken from Mathworks (2021)

Figure 1.5 shows the example of a typical CNN architecture. In addition to the convolutional layers, a typical CNN also has pooling (sub-sampling) layers, activation layers and fully connected layers. A pooling layer is usually introduced between every pair of convolutional layers to downsample the input and also avoid overfitting. However, they are not used in GANs because the abrupt discarding of hidden units during downsampling is detrimental for learning. Instead, a series of convolutional layers are stacked with different strides in order for the model to learn sub-sampling by itself during training. The *non-linear* activation layers, namely, Rectified Linear Units (ReLU) and Leaky ReLUs (LReLU)[2] were previously introduced in traditional NNs to augment their capability of approximating complex functions. The ReLU activation

[2] As shall be seen in section 3.5.2, ReLU and LReLU are used in the generator and discriminator architectures respectively.

function and its derivative are defined below. It essentially discards the negative values from a resulting feature map by explicitly setting them to zero.

$$f(x) = \max(0, x) \qquad \text{and} \qquad \frac{\partial f}{\partial x} = f'(x) = 1 \quad \forall \quad x > 0 \tag{1.1}$$

It can also be seen that the derivative $f'(x) = 0 \quad \forall \quad x < 0$ and is undefined when $x = 0$. ReLU is the most common non-linear activation function used in almost every modern DL architecture. This is mainly because it solves the vanishing gradient problem and also makes the training considerably faster. However, it suffers from the "dying ReLU" problem that consequently led to the introduction of the LReLU activation function.

Note that a unit is said to be inactive when its gradient is 0. Since there is no backward gradient flow through such units, they become perpetually stuck in the "inactive" state and die eventually. This causes a problem when a large number of units are inactive, effectively decreasing the model's capacity to learn because the gradient flow is stopped. This is avoided by using LReLU activations, which have a small negative slope for $x < 0$.

Leaky ReLU (LReLU) is a slightly modified version of ReLU, which is obtained by allowing the function to output small negative values also. It is given by:

$$f(x) = \begin{cases} x & \text{if} \quad x > 0 \\ 0.01x & \text{otherwise} \end{cases} \tag{1.2}$$

While this mitigates the dying ReLU problem, this modification comes at the cost of a slight decrease in performance compared to the original ReLU function.

1.5 Generative Adversarial Networks

There has been an exponential rise in the usage of GANs (Goodfellow *et al.*, 2014) for a wide variety of applications since their inception. Section 1.5.1 gives a thorough mathematical review of GANs.

1.5.1 Overview of GANs

A standard GAN consists of two neural networks (typically CNNs) called the generator (G), and the discriminator (D). The generator's task is to generate realistic images from input noise and try to fool the discriminator into classifying it is as a real image, whereas, the discriminator's task is to correctly classify the real and synthetic images given a single input. In other words, the discriminator performs a binary classification between a real input coming from the true data distribution and a synthetic input coming from the data distribution synthesized by the generator. The loss function is designed such that the generator is penalized if the discriminator is able to classify correctly (that is, reject the synthetic samples), so the feedback from the discriminator helps the generator update its weights and therefore, generate realistic images. At the same time, the discriminator is also penalized if it is unable to distinguish between the real and synthetic images. This type of competition, also known as the *minimax* game between the two networks results in the final output being realistic despite being a synthetic image. Figure 1.6 shows the basic GAN model. G generates a fake (synthetic) image x_g from the input noise[3] z sampled from a Gaussian distribution $p(z)$. Then, D performs a binary classification between the real image x_r sampled from the true distribution $p_r(x)$ and fake image x_g sampled from the synthetic distribution $p_g(x)$.

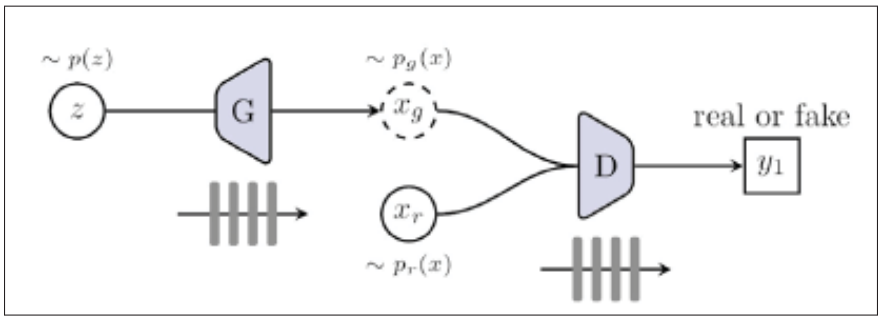


Figure 1.6 The basic GAN model
Taken from Yi *et al.* (2019)

[3] Note that the input can be a random noise image in the case of a CNN or a random noise vector in the case of a traditional NN.

The discriminator (D in figure 1.6) in a GAN is a simple binary classifier. The input for this network comes from two sources: (i) the real data instances, which it uses as positive examples, and (ii) synthetic instances from the generator, which it uses as negative examples.

The generator (G in figure 1.6) generates synthetic images by incorporating feedback from the discriminator. By giving noise as the input, the GAN learns to generate meaningful data by itself and also produces a wide variety of data that are similar to the target distribution. D plays a crucial role in training G because G's output does not affect the loss function directly (as will be seen in the next sub-section). Instead, its output is passed through D to get a "real" or "synthetic" (or, fake) label. This means that in order to obtain realistic samples, reliable gradient updates for the generator must be propagated through the discriminator during training.

1.5.2 Loss Functions in the Basic GAN

Since the objective of GANs is to mimic the distribution of the target data, the essence of the loss function is to reflect the distance between the real data distribution and the synthetic data distribution. The individual loss functions of the generator and discriminator are given by the following equations:

$$L_D^{GAN} = \max_D \left(\mathbb{E}_{x_r \sim p_r(x)} \left[\log(D(x_r) \right] + \mathbb{E}_{x_g \sim p_g(x)} \left[\log(1 - D(x_g) \right] \right) \tag{1.3}$$

$$L_G^{GAN} = \min_G \left(\mathbb{E}_{x_g \sim p_g(x)} \left[\log(1 - D(x_g) \right] \right) \quad \left(\text{where}, x_g = G_{\theta_g}(z) \right) \tag{1.4}$$

The combined loss function, also called as the *minimax* optimization is given by:

$$L_{G,D}^{GAN} = \min_G \max_D \left(\mathbb{E}_{x_r \sim p_r(x)} \left[\log(D(x_r) \right] + \mathbb{E}_{x_g \sim p_g(x)} \left[\log(1 - D(x_g) \right] \right) \tag{1.5}$$

where,

- $\mathbb{E}_{x_r \sim p_r(x)}$ is the expectation (average) over all "real" data instances x_r sampled from the true data distribution $p_r(x)$.
- $D(x_r)$ is the discriminator's probability estimate that a real data instance is indeed "real".

- $\mathbb{E}_{x_g \sim p_g(x)}$ is the expectation over all generated instances x_g sampled from the synthetic data distribution $p_g(x)$.
- $D(x_g)$ is the discriminator's estimate of the probability that a synthetic data instance is real. x_g is also written as $G_{\theta_g}(z)$, which denotes a non-linear mapping from a vector/image in z-space to x_g parameterized by θ_g.

As can be seen in equation 1.3, D is a simply binary classifier with a maximum likelihood objective. Equation 1.5 is effectively a zero-sum non-cooperative game where the either the discriminator D or the generator G wins at the expense of the other. An intuitive way of understanding the GAN dynamics (equation 1.5) is as follows: D tries to maximize each of the two terms, that is, by classifying the real samples correctly ($D(x_r)$ is high) it maximizes the first term, whereas, by rejecting the fake samples ($D(x_g)$ is low) it tries to maximize the second term. On the other hand, G is trying to minimize the objective by fooling the discriminator (and generating realistic samples) such that $D(x_g)$ is high.

In theory, the convergence for such an optimization objective is given by the Nash equilibrium where the generator's samples are indistinguishable from the true data while the discriminator only outputs a constant value of 0.5 irrespective of the type of input. However, in practice, due to the very nature of their construction, it is difficult to achieve convergence in GAN training. This is because, in a minimax game, when the objective for one player is to maximize the loss function, while the other's is to minimize it, the stationary point for both players tends to be a *saddle point*, which is neither a local minimum or local maximum.

CHAPTER 2

LITERATURE REVIEW

This chapter introduces the literature related to the vertebral bone segmentation problem. It covers the major advancements but also highlights the drawbacks of the existing methods, thereby motivating the need for the proposed solution.

Vertebral bone segmentation in MR images is a well-studied problem, with approaches that can be classified into semi-automatic (section 2.1), graph-based (section 2.2), and learning-based (section 2.3). The types of image synthesis methods are discussed in section 2.4.1 and the recently proposed GAN-based medical image synthesis methods are covered in section 2.4.2 to highlight the practicality of this approach. However, there is a general lack of learning-based methods for the specific problem of segmenting scoliotic spines. This is primarily due to the unavailability of large and labeled spine datasets, let alone scoliotic spine datasets, for training. Section 2.5 discusses the literature on Bayesian uncertainty estimation.

2.1 Semi-automatic Approaches

Semi-automatic approaches rely on user-interaction for the initialization of seed points on the vertebrae of interest. Based on a single point-in-vertebra initialization, Zukic et al. (2012) proposed a segmentation method in MR images using multiple-feature boundary classification and iterative mesh inflation. The initialized seed points are used by the iterative closest point algorithm (ICP) (Chen & Medioni (1992), Besl & McKay (1992)) for determining a rough orientation of the vertebra and the vertebral boundaries are classified based on the voxel probabilities estimated from edge and intensity based features. Post vertebral boundary classification, the segmentation is achieved by iteratively inflating a mesh using ballon forces towards star-shaped geometry and enforcing mesh smoothness using butterfly subdivision hierarchy scheme. A total of 11 spine datasets containing 92 vertebrae were used, which were manually segmented in the sagittal plane by expert neurosurgeons. On comparing their method's performance with manual and automatic seed point initialization using quantitative metrics such

as the Dice Similarity Coefficient and mean distance error between the reference and segmented surfaces, it was noted that manually initializing the seed points for all the vertebrae and the background achieved the best result. Suzani et al. (2014) proposed a method for segmenting the lumbar (L1-L5) vertebral bodies from MR images using the expectation maximization (EM) algorithm to align a statistical model from manually segmented CT volumes and the user-initialized Canny edge-detected vertebrae. The MR images are pre-processed for bias field correction and smoothed using 3D anisotroic diffusion filtering. A user is then tasked with initializing five points in the lumbar vertebrae for the Canny edge detection algorithm. A statistical multi-vertebrae, shape and pose model from their previous work (Rasoulian, Rohling & Abolmaesumi, 2013) and the edge-detected vertebrae are registered using the EM algorithm. An important note is that seed initializations are done on a per-slice basis. A dataset of 9 patients along with their ground truth segmentations were used for quantitative evaluation using the Hausdorff distance.

A major drawback of these methods is that the quality of segmentation depends on how well the points are initialized. Therefore, the onus is on the end-user to initialize the points accurately. The failure cases caused by issues with initialization are generally not discussed. In this regard, our primary motivation is to develop an automatic method that involves the end-user as little as possible.

2.2 Graph-based Segmentation Methods

Egger et al. (2012) described a graph-based method that performed rectangle-shaped graph-cuts for segmenting vertebrae from 2D MR images. A similar study by Schwarzenberg et al. (2014) extends the previous approach to 3D by using a cubic-shaped distribution of the graph's nodes for volumetric segmentation. A directed graph is set up based on a single user-defined a seed-point inside the object to be segmented. A closed contour is formed by sending out rays radially from the seed point to obtain the intersection with the object boundary.

In graph-based methods, the distinction between foreground and background is important and the user is responsible for labeling/initializing the seed points such that the algorithms do not segment unwanted regions. Similar to the semi-automatic approaches, such segmentation methods cannot be easily scaled to a large number of volumes because some user interaction is required.

2.3 Learning-based Methods

Huang et al. (2009) proposed an automatic segmentation method consisting of three stages: a modified AdaBoost algorithm for vertebral candidate detection trained on feature vectors obtained by an overcomplete Haar wavelet transformation followed by a refinement step via robust curve fitting for eliminating false positive and recovering missed vertebral regions, and finally, vertebrae segmentation using an iterative normalized-cut algorithm. The detected vertebrae act as seed points for an iterative segmentation of the vertebrae, making it a fully automatic method. The training dataset consisted of 22 spinal MRI datasets containing the entire vertebral column. A total of 1398 images were manually labeled for training the AdaBoost algorithm, which constituted the set of positive samples. Additionally, vertebrae were marked manually pixel-wise for quantitative assessment. While the method proposed is automatic, the reported results only included the segmentation of the vertebral bodies in sagittal MR slices. This leaves out the spinous processes, which is important for estimating the orientation of the vertebral bodies in 3D spine volumes.

Chu et al. (2015) addressed the problem of localization and segmentation of vertebral bodies in both 3D MR and CT images by using a unified random forest regression and classification framework. The target vertebral body is localized by aggregating votes from randomly sampled 3D image patches to obtain a probability map by training a random forest regressor with manually delineated boundaries. The localized vertebral body defines a region of interest (ROI) that is utilized further for two tasks: (i) estimating the likelihood of a voxel being in the foreground or background using random forest soft classification, and (ii) estimating a spatial prior map. The voxel likelihood and the prior probability map are then combined to calculate the posterior

probability map for each voxel in the ROI. The final segmentation is achieved by using binary thresholding that keeps retains the largest connected component. A dataset containing 23 3D MR images of the lower spine (T11-L5) and 10 3D spine CT images was used. In both datasets, vertebral bodies were manually identified and segmented. Due to the availability of ground truths, quantitative evaluation included the computation of Dice coefficients, the average absolute and Hausdorff distances. While the results included 3D segmentations, they suffer from the same issues as those of Huang *et al.* (2009), in that only the vertebral bodies are segmented in sagittal slices leaving out the spinous processes and axial/coronal segmentations. For our problem, these structures are necessary for reconstructing a complete 3D model of the spine.

Neubert et al. (2012) and Guerroumi et al. (2019) proposed automatic methods for simultaneously segmenting the vertebral bodies and interverterbral discs. While the former performed a 3D segmentation of the thoracic and lumbar spine using statistical shape analysis models, the latter proposed a bone segmentation method of the thoracic spine from 2D MR images by incorporating a squeeze-and-excitation block in a U-Net-based network architecture. It is important to note that Guerroumi *et al.* (2019)'s work is the first approach that performs segmentation on scoliotic spine images. They also mentioned manually labelling the vertebral bodies in 1152 2D slices.

A common observation across the literature is that most methods used manual segmentations as the ground truth, with the main regions of interest being the vertebral bodies. While this may (arguably) be less painstaking when each spine volume has fewer slices (~ 20), it quickly turns into a major hindrance when each volume has about 216 2D slices on average and manually segmenting each slice is impractical. It also does not scale well to the case where one has hundreds of training volumes. This, therefore, provides strong motivation for an approach based on unsupervised learning. One should also note that there is a clear dearth of literature pertaining to the segmentation of scoliotic MR spines. The nature of the disease and the general unavailability of training datasets brings complexity to the problem, and makes segmentation a challenging task.

2.4 Image Synthesis Methods

This section discusses the types of image synthesis methods using GANs (section 2.4.1) and the distinction between paired and unpaired data. The literature concerning GANs for the task of medical image synthesis is covered in section 2.4.2.

2.4.1 Types of Image Synthesis Methods

There are two particular reasons why GANs are popular for image synthesis (Kazeminia et al., 2019):

1. They discover the high dimensional *latent* distribution of the data **without** having to explicitly model the true data distribution and **without** using any *approximate inference* methods,
2. They are able to maximize the probability density over the data generating distribution.

One of the long-standing problems in medical image analysis is the general lack of labeled data and in some cases unavailability of sufficient training data for training DL models. Therefore, it is difficult to implement supervised learning strategies in such cases. There are two ways to artificially synthesize medical images (Yi *et al.*, 2019):

- **Unconditional synthesis**: The images are generated from random noise without any other input to or, *conditioning* on, the network. An example of this method is the working of the original GAN as explained in section 1.5.1.
- **Cross modality synthesis** (CMS): The idea of generating images from one imaging modality to another by keeping the underlying style and structure of the original modality intact. This can be further divided into two categories: supervised and unsupervised CMS.

A GAN could be made to generate images with specific properties based on some additional information provided to the generator and discriminator. This "additional" information is called *conditioning*[1] and hence the name conditional GANs (cGANs). The framework which first proposed this idea is called pix2pix (Isola et al., 2017) and it comes under the category

[1] It is useful to note that the term comes from probability theory. Given two random variables X and Y, we can provide more information about X by *conditioning* it upon Y i.e. $X|Y$, read as X "given" Y.

of supervised CMS. The most common conditioning inputs are ground truth images whose distribution we want to mimic. Other types of conditioning data include class labels, spatial coordinates of objects, text descriptions, etc. The rest of this thesis is based on the second category of CMS, which was first proposed by Zhu et al. (2017) using the CycleGAN framework. This is elaborately discussed in chapter 3.

Paired Data vs. Unpaired Data

The CycleGAN model is instrumental in understanding the popularity of image-to-image translation and its eventual success in medical image synthesis. It is different from most of the GAN architectures in that it uses *unpaired* images for training. For the work described in this thesis, this means that the model does not require paired MR and CT images of a particular patient, instead, it is trained with MR images of one set of patients and CT images of another set of patients with the datasets being *mutually exclusive*. However, to ensure good results, the underlying anatomy being imaged must be the same (for instance, brain-brain, spine-spine, etc.). This is especially useful in cross-modality medical image synthesis because it not always ideal to acquire the anatomical data of the same patient using different imaging modalities. It is time-consuming and particularly hazardous in some cases. Therefore, despite the lack of correspondence between the images, the ability of CycleGANs to translate images with high realism is what makes them powerful for medical images synthesis also.

2.4.2 GAN-based Medical Image Synthesis

This section reviews some of the recent literature that used GANs for medical image synthesis. Keeping in mind that GANs were originally developed for artificially generating natural images, one can likewise find several applications of GANs for computer vision tasks. We found that the GAN-based literature in medical image synthesis mostly focuses on synthesizing anatomical regions of the body other than the spine across different imaging modalities. However, they present an important step towards the proliferation of unsupervised cross-modality synthesis for medical images. To the best of our knowledge, there has not been any study yet that uses a

GAN-based volume-translation approach for facilitating the 3D segmentation of scoliotic spines. The novelty here lies in the fact that this approach is used on the scoliotic spines, which present two important challenges: (i) the general lack of scoliotic data, and (ii) complex curvatures of the spine.

Wolterink et al. (2017) used the standard CycleGAN model for unpaired MR-CT synthesis of brain images. Though the training is done on unpaired data, the quantitative results are evaluated using supervised metrics, namely, the mean absolute error (MAE) and peak signal-to-noise ration (PSNR). The synthetic and reference CT images of the same patient were pre-aligned using rigid registration based on mutual information. This is the standard method for evaluating synthesized images, however, it cannot be used in the absence of paired data.

Modified versions of CycleGAN have also been proposed to better suit the features of medical images. Specifically, Hiasa et al. (2018) used gradient-consistency loss to improve the simulated image quality at the boundaries for their dataset containing MR and CT images of the pelvic region. While the evaluation criteria were the same as in Wolterink *et al.* (2017) where MR and CT images were pre-registered, they used mutual information (MI) to compare the real and synthesized images across different domains. Armanious et al. (2019) proposed the use of perceptual and style losses (Johnson et al., 2016) for PET-CT translation and MR motion correction. Their incentive to adapt perceptual and style losses was motivated from the fact that direct pixel-wise losses failed to capture the perceptual quality of the images, in general. Therefore, by minimizing the perceptual discrepancies, global consistency can be enhanced in the translated images. However, this also comes at the high cost of training a separate CNN for calculating the style and perceptual losses, in addition to the CycleGAN.

Zhang et al. (2018) augmented the standard CycleGAN by incorporating a shape-consistency loss to constrain the geometric invariance of the synthetic data and performed volume-to-volume translation for segmenting 3D cardiovascular MR and CT data. The goal of the shape-consistency loss is to restrict the geometric variations possible, by projecting the translated data into a shared

latent space and computing the pixel-wise semantic relationship. Theirs is the first approach that uses a fully 3D method for segmenting multi-modal volumes.

It is important to note that while the modifications to the original CycleGAN proposed in the literature improve its performance on medical images, they typically comes at the cost of training at least one more network. We shall see in chapter 3 that the original CycleGAN model contains 4 networks that are trained simultaneously. Considering the three-dimensional nature of our problem, this already reaches the memory limit of most high-end GPUs. Therefore, training additional networks on top of the CycleGAN model becomes extremely difficult as one has to work within the constraints of the memory limits.

2.5 Uncertainty Estimation

Quantifying what a DL model does not know is crucial in high-stakes application domains such as medical imaging. The point-estimate outputs (classification accuracies, for instance) are blindly believed to be a true representation of the model's confidence, which is not always the case as models with high predictive probabilities (the softmax outputs) can also be uncertain (Gal & Ghahramani, 2016). With an estimate of model uncertainty, we can expect the model to output high uncertainty when out-of-distribution data are seen. Such outputs could then be passed through an expert for further supervision, thereby preventing erroneous decisions. Bayesian probability theory offers mathematically principled methods for quantifying uncertainty in theory, however, they cannot be realized in practice due to their intractability. On the other hand, approximate inference methods can have recently gained importance due to their superior ability in tractably approximating the true posterior probability distributions.

Gal & Ghahramani (2016) proposed *Monte Carlo (MC) Dropout* - an incremental method that uses standard dropout (Srivastava et al., 2014) in NNs as an approximate Bayesian inference to Gaussian processes[2] (Rasmussen & Williams, 2005). At its heart, MC Dropout is a simple

[2] Gaussian processses (GPs) are well known probabilistic models that readily provide uncertainty estimates in their predictions by the virtue of their mathematical construction. However, they cannot be trivially scaled to large dimensional inputs on the order of magnitude that is common in deep NNs.

computation added to the standard training regime of NNs trained with dropout. Instead of disabling dropout during test-time, each input is passed through the network with dropout enabled for a fixed number of times (known as the Monte Carlo samples), such that each stochastic forward pass results in a new set of weights being utilized for the computation of the final output. Thus, a distribution is generated over the output space for each new test input. The mean and variance of the MC samples are then called the predictive mean and variance, respectively, and provide an estimate of model uncertainty. In theory, it is shown mathematically that the dropout training objective essentially minimizes the Kullback-Liebler (KL) divergence between the approximate (MC dropout) distribution and the posterior of a deep GP.

Kendall & Gal (2017) built upon the work of Gal & Ghahramani (2016) and combined the two types of uncertainties: *aleatoric* and *epistemic*, in the context of (supervised) semantic segmentation and depth regression tasks. *Aleatoric* uncertainty captures the noise inherent in the data, such as the noise accumulated due to sensor errors or the noise arising due to varying consensus of manually segmented labels by experts. Irrespective of the data collected, these errors cannot be reduced. On the other hand, *epistemic* uncertainty (also known as model uncertainty (Gal & Ghahramani, 2016)) captures our ignorance about which particular model represents the given data. In other words, it captures the uncertainty of the model over its parameters and the architecture. Unlike the irreducible aleatoric uncertainty, it can be reduced to zero (or, explained away) in the limit of infinite data. Their main contribution was the unification of both uncertainties in a single DL model and it was shown that estimating aleatoric uncertainty typically improves the models' performance over their non-Bayesian counterparts. The technicalities of how this unification is achieved are presented in chapter 3, section 3.4.3.

In one of the few works that discuss the importance of uncertainty quantification in medical imaging, Hemsley et al. (2020) estimated the aleatoric and epistemic uncertainties in the MR-CT synthesis of the brain using conditional GANs. Their method was primarily inspired by the work of Kendall & Gal (2017). The L_1-norm loss function in the conditional GAN (Isola *et al.*, 2017) was modified such that the model also learned to estimate aleatoric uncertainty and the epistemic

uncertainty was calculated using MC Dropout. It should be noted that theirs is the only work that proposed the idea of uncertainty estimation in (supervised) medical image synthesis.

Considering our primary goal of 3D segmentation and volume reconstruction of the vertebral column, there are some major drawbacks and gaps in the literature:

1. Most methods use 2D images due to their ease of usage and low computational requirement, but, they fail to capture the spatial correlation between slices (for e.g. the degree of curvature), which is crucial (Zukic *et al.*, 2012). Therefore, in such cases, 3D methods are essential.
2. Human supervision, either in terms of manual segmentation of the ground truth labels or via user-interaction, is laborious and prevents scalability to a large number of volumes, and
3. A general lack of segmentation methods involving scoliotic spinal images.
4. Due to the unsupervised nature of the problem, no methods exist in the literature that discuss the quantification of epistemic and aleatoric uncertainties in unsupervised medical image synthesis.

Therefore, our contribution in the form of the proposed solutions in this thesis are *two-fold*:

- The first three issues are addressed by performing a volume-to-volume translation for facilitating the segmentation of scoliotic spines and reconstructing their 3D models, *all without the need for labeled training data*. The advantages of our method lie in the fact that the training is *completely unsupervised* and works with *unpaired data*.
- In order to make model more self-sufficient and address the fourth issue above, the aleatoric and epistemic uncertainties are estimated by proposing a Bayesian adaptation of the 3D CycleGAN model.

CHAPTER 3

METHODOLOGY

As stated in chapter 2, the CycleGAN framework lies at the heart of our method. This chapter mainly focuses on a thorough review of the working of the CycleGAN for volume-to-volume translation, the segmentation and reconstruction steps that follow, and the mathematical formulation of aleatoric and epistemic uncertainties within the proposed CycleGAN framework. The structure of this chapter is as follows: section 3.1 gives an overview of the proposed three-stage method and describes the CycleGAN model, section 3.2 discusses the idea of unsupervised translation and the specific loss functions used, section 3.3 presents the Otsu segmentation and reconstruction method, section 3.4 discusses uncertainty estimation and the unification of both uncertainties. Section 3.5 presents the experimental protocal which is further divided into sections 3.5.1 and 3.5.2 that describe the dataset used and provide details on training the 3D model, respectively. Finally, section 3.5.3 describes the approach taken for quantitative validation of the segmented vertebrae in detail.

3.1 CycleGAN Overview

The proposed segmentation method consists of three stages as shown in figure 3.1. We train a 3D CycleGAN model with gradient consistency loss from scratch to perform unsupervised volume-to-volume translation from MR to CT domains. The Otsu thresholding algorithm is then applied to the resulting synthesized CT volumes to easily segment the bones. Finally, a complete 3D model of the scoliotic spine is reconstructed based on the resulting segmentation.

Volume-to-volume translation is essentially a 3D extension of image-to-image translation to volumetric data. While processing images is less computationally intensive than processing volumes, learning based on 2D images obtained by slicing the volumes does not account for the correlation between the slices and essential 3D information is lost. Figure 3.2 shows the resulting 2D image-to-image translation between three *consecutive* sagittal MR slices. The red and blue arrows show how the shapes of the spinous processes are completely missed. Therefore,

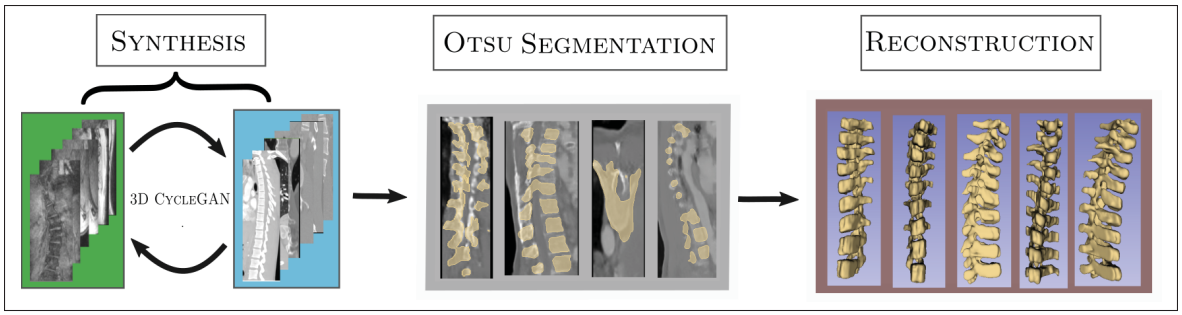


Figure 3.1 The three stages of our proposed solution: Synthesis, Otsu Segmentation and 3D Model Reconstruction

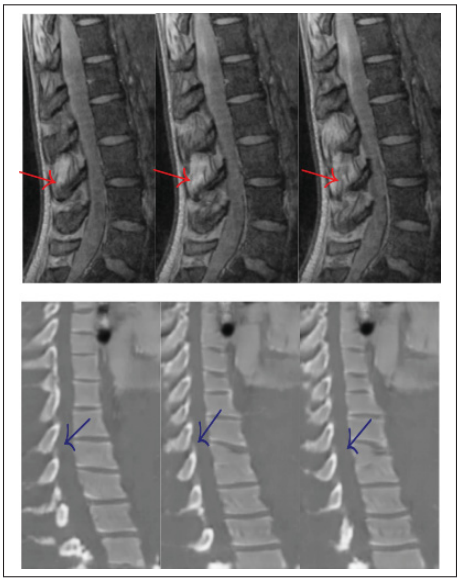


Figure 3.2 *Top Row:* Three consecutive images sliced in the sagittal axis from a MR volume. *Bottom Row:* Corresponding translated CT slices. Red and Blue arrows shows the missed spatial information

for a good segmentation, we found that it is best for the model to learn the correlation between the slices by itself. Hence, by using the 3D CycleGAN model, we proposed to work directly on the volumetric data. The main idea is as follows: given a set of unpaired volumes from MR and CT domains, the model learns two function mappings simultaneously using two generators $G_{\text{MR}\rightarrow\text{CT}}$ and $G_{\text{CT}\rightarrow\text{MR}}$. Since voxel-wise comparison is infeasible due to the unavailability of paired data, the *cycle-consistency loss* (Zhu *et al.*, 2017) is introduced as a comparison metric.

The idea is straightforward: an input MR volume translated to the CT domain should ideally be recovered when that synthesized CT volume is translated back to the MR domain and vice-versa, that is, $I_{\text{MR}} \approx G_{\text{CT}\to\text{MR}}(G_{\text{MR}\to\text{CT}}(I_{\text{MR}}))$ and $I_{\text{CT}} \approx G_{\text{MR}\to\text{CT}}(G_{\text{CT}\to\text{MR}}(I_{\text{CT}}))$. In order for this reformulation to give reliable learning information, two discriminators D_{MR} and D_{CT} are used, whose task is to distinguish between the real volumes $(I_{\text{MR}}, I_{\text{CT}})$ and simulated volumes $(G_{\text{CT}\to\text{MR}}(I_{\text{CT}}), G_{\text{MR}\to\text{CT}}(I_{\text{MR}}))$ respectively.

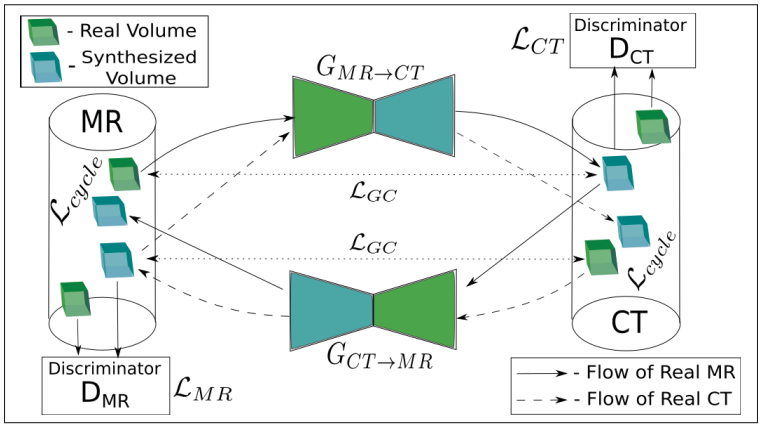


Figure 3.3 Working of the CycleGAN model
Figure adapted from Hiasa *et al.* (2018)

Figure 3.3 illustrates the main idea. Each cylinder represents an imaging modality with green boxes showing the real volumes and blue boxes showing the synthetic volumes of that domain. There are two *cycles* within the model:

- **Forward Cycle**: (shown by solid arrows) Starting from green box (left, top), going through $G_{\text{MR}\to\text{CT}}$ to blue box (right, top), then going through $G_{\text{CT}\to\text{MR}}$ to recover blue box (left, top).
- **Backward Cycle**: (shown by dashed arrows) Starting from green box (right, bottom), going through $G_{\text{CT}\to\text{MR}}$ to blue box (left, bottom), then going through $G_{\text{MR}\to\text{CT}}$ to recover blue box (right, bottom).

The cycle-consistency loss ($\mathcal{L}_{\text{Cycle}}$ & section 3.2.2) is calculated between the real and recovered volumes (top left and bottom right in fig. 3.3). The gradient-consistency loss ($\mathcal{L}_{\text{GC}}$ & section 3.2.3) is calculated between the real and synthesized volumes (top left, top right and bottom left, bottom right in fig. 3.3).

3.2 Unsupervised Volume-to-Volume Translation

The cycle-consistency loss is combined with the adversarial and gradient consistency losses for performing the unsupervised volume-to-volume translation. Therefore, our model consists of three types of loss terms in total: the adversarial losses ($\mathcal{L}_{\text{CT}}$ and $\mathcal{L}_{\text{MR}}$), the cycle-consistency loss ($\mathcal{L}_{\text{Cycle}}$) and the gradient consistency loss ($\mathcal{L}_{\text{GC}}$). We show the full optimization objective first, followed by a brief description of each loss term. The optimization objective is defined as:

$$\left(G^*_{\text{CT}\to\text{MR}}, G^*_{\text{MR}\to\text{CT}}\right) = \arg \min_{\substack{G_{\text{MR}\to\text{CT}} \\ G_{\text{CT}\to\text{MR}}}} \max_{\substack{D_{\text{MR}} \\ D_{\text{CT}}}} (\mathcal{L}_{\text{CT}} + \mathcal{L}_{\text{MR}} + \lambda\mathcal{L}_{\text{Cycle}} + \gamma\mathcal{L}_{\text{GC}}), \tag{3.1}$$

where λ and γ are the hyperparameters for weighting cycle- and gradient-consistency losses. We set $\lambda = 10.0$ and $\gamma = 0.1\lambda$ for this study (obtained empirically using hints from (Zhu *et al.*, 2017) and (Hiasa *et al.*, 2018)).

3.2.1 Adversarial Loss

The adversarial loss helps in mapping the source data distribution to the target data distribution. For the mapping defined by $G_{\text{MR}\to\text{CT}} : I_{\text{MR}} \to I_{\text{CT}}$ and its discriminator D_{CT}, the objective is defined as:

$$\mathcal{L}_{\text{CT}} = \mathbb{E}_{x\sim I_{\text{CT}}}[\log D_{\text{CT}}(x)] + \mathbb{E}_{y\sim I_{\text{MR}}}[\log(1 - D_{\text{CT}}(G_{\text{MR}\to\text{CT}}(y)))]. \tag{3.2}$$

Similarly, the objective for the reverse path is defined as:

$$\mathcal{L}_{\text{MR}} = \mathbb{E}_{y\sim I_{\text{MR}}}[\log D_{\text{MR}}(y)] + \mathbb{E}_{x\sim I_{\text{CT}}}[\log(1 - D_{\text{MR}}(G_{\text{CT}\to\text{MR}}(x)))], \tag{3.3}$$

where x and y are the volumes from the CT and MR domains respectively. The reader is referred to figure 3.3 for visual illustration. In equation 3.2, $D_{\text{CT}}(x)$ is the probability assigned by the discriminator for a real CT volume. Likewise, $D_{\text{CT}}(G_{\text{MR}\to\text{CT}}(y))$ is the probability assigned to a synthetic CT volume translated from the MR domain. The goal of the discriminator D_{CT} is to

maximize $\mathcal{L}_{\text{CT}}$ such that $D_{\text{CT}}(x)$ is high and $D_{\text{CT}}(G_{\text{MR}\to\text{CT}}(y))$ is low, while the generator's $G_{\text{MR}\to\text{CT}}$ objective is to minimize $\mathcal{L}_{\text{CT}}$ such that $D_{\text{CT}}(G_{\text{MR}\to\text{CT}}(y))$ is high. The training dynamics are identical for the backward cycle concerning $\mathcal{L}_{\text{MR}}$.

3.2.2 Cycle-consistency Loss

The generators are enforced to be cycle-consistent in order to learn the function mappings with *unpaired* data. In other words, given an input MR volume, the generators should be able to recover it after consecutive forward and backward translations, with as much similarity as possible i.e. $y \approx G_{CT\to MR}(G_{MR\to CT}(y))$ and vice-versa. To steer the learning towards this constraint, the cycle-consistency loss is defined as:

$$\mathcal{L}_{\text{Cycle}} = \mathbb{E}_{x\sim I_{\text{CT}}}[||G_{\text{MR}\to\text{CT}}(G_{\text{CT}\to\text{MR}}(x)) - x||_1] + \mathbb{E}_{y\sim I_{\text{MR}}}[||G_{\text{CT}\to\text{MR}}(G_{\text{MR}\to\text{CT}}(y)) - y||_1], \tag{3.4}$$

where $||\cdot||_1$ denotes the L_1 norm between the real and recovered volumes for each domain.

3.2.3 Gradient-consistency Loss

Since our objective is to use the synthesized CT volumes for facilitating the segmentation of the bones, we place more emphasis on translating the bone boundaries as accurately as possible, as opposed to their insides. Therefore, a commonly used image-similarity metric called gradient correlation is utilized as an additional constraint. It is defined as the normalized cross correlation (NCC) between the gradients of two images (see equation 3.6). This is because gradient measures help in filtering out the low frequency differences between images that are caused by soft tissues (Penney et al., 1998). Moreover, given the difference in the voxel intensities of bones in MR and CT volumes, maximizing cross correlation between them focuses on registering the bone anatomy that is prominent in CT volumes due to the inherent nature of image-gradients acting as a high pass filter. Gradient correlation similarity has shown promising results in the context of vertebral level localization (Silva et al., 2016). The gradient correlation between two volumes

is defined as:

$$GC(A,B) = \frac{1}{2}\left(NCC(\nabla_X A, \nabla_X B) + NCC(\nabla_Y A, \nabla_Y B) + NCC(\nabla_Z A, \nabla_Z B)\right) \tag{3.5}$$

$$\text{where,} \quad NCC(\nabla A, \nabla B) = \left(\frac{\sum_{i,j}(\nabla A - \mu_{\nabla A})(\nabla B - \mu_{\nabla B})}{\sqrt{\sum_{i,j}(\nabla A - \mu_{\nabla A})^2}\sqrt{\sum_{i,j}(\nabla B - \mu_{\nabla B})^2}}\right) \tag{3.6}$$

where, ∇ represents the gradients of the input volume in X, Y and Z directions respectively. $\mu_{\nabla J}$ is the mean of the gradient of volume J. Therefore, gradient consistency (GC) loss is defined as:

$$\mathcal{L}_{\text{GC}} = \frac{1}{2}\left[\mathbb{E}_{x\sim I_{\text{CT}}}(1 - GC(x, G_{\text{CT}\to\text{MR}}(x))) + \mathbb{E}_{y\sim I_{\text{MR}}}(1 - GC(y, G_{\text{MR}\to\text{CT}}(y)))\right] \tag{3.7}$$

3.3 Otsu Thresholding and Volume Reconstruction

Otsu's method (Otsu, 1979) is a widely used image thresholding algorithm that maximizes inter-class pixel intensity variance and returns a single intensity threshold, thus separating the image into foreground and background pixels. In other words, given a grayscale image, it finds an intensity threshold value that minimizes the intra-class variance by minimizing the weighted sum of the variances of the two classes. We have the following variance equation:

$$\sigma_\omega^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t) \tag{3.8}$$

where, ω_0, ω_1 are the probabilities of the classes being separated by a threshold t, and σ_0^2, σ_1^2 are the variances of the two classes. The class probabilities are calculated by initially dividing the image into L bins of the histogram.

$$\omega_0(t) = \sum_{i=0}^{t-1} p(i) \qquad \omega_1(t) = \sum_{i=t}^{L-1} p(i)$$

where, $p(i)$ is the number of pixels with intesity i.

We used the Otsu thresholding option provided in the “Segment Editor” module by 3D Slicer[1] Since the training was done on unpaired data, a few segmentation errors were observed. We used 3D Slicer’s "Paint Brush" tool for hole-filling and also performed median smoothing operations to reduce noise while preserving the edges. The resulting post-processed segmentation was used for generating a 3D model of the scoliotic spine provided in the Segment Editor module. We emphasize the relative ease with which the bones can be segmented using our framework. The resulting outputs at each stage of our method are shown in section 4.2.

3.4 Uncertainty Quantification

The section presents the mathematical details underlying the estimation of epistemic and aleatoric uncertainties (sections 3.4.1 and 3.4.2, respectively). The corresponding results and discussion are presented in chapter 4.

3.4.1 Epistemic Uncertainty

Epistemic uncertainty refers to the uncertainty in the model parameters. In other words, since the neural networks (NNs) are function mappers that try to best represent the data under the (user-defined) architectural constraints, it captures the uncertainty arising from which particular model generated the data. Therefore, it is also known as model-dependent uncertainty. In theory, it is straightforward to calculate epistemic uncertainty - a Bayesian NN is trained by replacing the deterministic weight parameters by a distribution over each weight and the posterior distribution over the weights $p(W|X,Y)$ is the computed i.e. all the sets of weight parameters given the dataset. Let us consider its expansion using Bayes’ theorem:

$$p(W|X,Y) = \frac{p(Y|X,W)p(W)}{p(Y|X)} = \frac{p(Y|X,W)p(W)}{\int_W p(Y|X,W)p(W)dW}$$

In practice, the posterior $p(W|X,Y)$ is difficult to compute because of the marginal probability $p(Y|X)$ in the denominator of the equation above. The integral is computed over *all* the possible

[1] https://www.slicer.org/

weight parameters, which is intractable. Therefore, we turn towards approximate inference methods to efficiently compute the posterior.

Monte Carlo (MC) dropout (Gal & Ghahramani, 2016) is used in this study, where training a NN with dropout effectively minimizes the KL divergence between an approximate distribution and the true posterior of a deep GP. Mathematically, this is interpreted as $\mathrm{KL}(q^*_\theta(W)||p(W|X,Y))$, where $q^*_\theta(W)$ (parameterized by θ) is a simple variational distribution and $p(W|X,Y)$ is the true model posterior. The crux of this method lies in the reformulation of the standard dropout as a Bayesian approximation, wherein, the approximating distribution is a bimodal mixture of Gaussians with one component having non-zero mean with probability p_i and the other having a zero mean with probability $(1 - p_i)$. In addition, both these components have extremely small variances. One should observe that this reformulation is indeed another way of representing the Bernoulli distribution that randomly sets weights to zeroes with probability p_i, but with Gaussian components.

In practice, the network is trained with dropout applied before every weight layer. During inference, T stochastic forward passes are performed through the network with dropout enabled (where, T refers to the number of MC samples). With each stochastic forward pass, the weights $\{\hat{w}\}_{t=1}^{T}$ are sampled from the approximate posterior $q^*_\theta(W)$ i.e. $\hat{w} \sim q^*_\theta(W)$ resulting in an unbiased estimate of the prediction $\{\hat{y}^{\hat{w}_t}\}_{t=1}^{T}$. Then, the mean and variance of T stochastic forward passes are computed, which are the predictive mean and model uncertainty (predictive variance), respectively. Mathematically, the predictive mean is $\mathbb{E}\left[\hat{y}^{\hat{w}_t}\right] = \frac{1}{T}\sum_{t=1}^{T}\hat{y}^{\hat{w}_t}$ and model uncertainty is $\mathrm{Var}\left(\hat{y}^{\hat{w}_t}\right)$.

3.4.2 Aleatoric Uncertainty

Aleatoric uncertainty measures the uncertainty that can be attributed to the data as a consequence of noise in the inputs. Therefore, it is also known as the data-dependent uncertainty. It can be further divided into two categories: (i) *homoscedastic* aleatoric uncertainty, and (ii) *heteroscedastic* aleatoric uncertainty. The former assumes identical observation noise for

every input, whereas the latter assumes that the noise can vary with each input. Modeling heteroscedastic uncertainty is useful in cases where one might anticipate higher noise in certain inputs in the data space compared to others (for e.g., data acquired using different parameter settings of the imaging modality, or labeling of the ground truths by various experts). In this thesis, since the data is acquired from different sources as mentioned in section 3.5.1, the heteroscedastic model is used.

Fundamentally, all image synthesis tasks can be interpreted as regression tasks, in that, the ultimate goal is to generate (*regress*) new pixels that have similar style and content characteristics of the source images. Capturing aleatoric uncertainty provides two advantages: (i) it helps in identifying the sources of noise in the input, and (ii) it acts a loss function attenuator (Kendall & Gal, 2017), which is explained later in the context of the cycle-consistency loss. In standard Bayesian NNs (BNNs), the Gaussian likelihood is used to model aleatoric uncertainty, which gives the following loss function:

$$\mathcal{L}_{BNN}(\theta) = \frac{1}{D}\sum_{i}\frac{||y_i - \hat{y}_i||^2}{2\hat{\sigma}_i^2} + \frac{1}{2}\log\left(\hat{\sigma}_i^2\right) \quad \text{(where, } \hat{y} = f_\theta(x)\text{)} \tag{3.9}$$

where, D is the total number of pixels, $\hat{y}_i$ is the model's prediction, y_i is the ground truth, and $\hat{\sigma}_i^2$ is the predicted variance for pixel i. The output $\hat{y}$ is obtained by passing an input x to the BNN model f, parameterized by the weight parameters θ. Note that there are two terms in the above loss function: (i) the standard *supervised* loss for regression, and (ii) the uncertainty regularization term. The uncertainty (variance, $\hat{\sigma}^2$) is learned by the model implicitly and does not require supervision in the form of uncertainty ground truth labels.

Considering the first term in equation 3.9, an issue arises when adapting it to the unsupervised learning problems. One must note that aleatoric uncertainty captures the noise inherent in the data, which manifests itself as the noise inherent in the ground truth data[2] (y_i). Therefore, it cannot be directly adapted to the CycleGAN model because the training data is unpaired and ground truth CT volumes are unavailable. In order to tackle this issue, this thesis proposes a

[2] In non-Bayesian NNs, this is generally overlooked as ground truth data is assumed to be accurate.

novel adaptation to the cycle-consistency loss that helps us extract the aleatoric uncertainty. Recall from section 3.2.2 and equation 3.4 that the cycle-consistency loss computes the L_1-norm between the recovered volume and the original volume. Therefore, it can be seen that the original volume acts as an indirect ground truth for the recovered volume so that the major characteristics of the original volume remain intact during translation. Hence, the aleatoric uncertainty in the 3D CycleGAN model is computed by the updating equation 3.4 in the following manner:

$$\mathcal{L}_{\text{AleaCycle}} = \left(\mathbb{E}_{x \sim I_{\text{CT}}} \left[\frac{||G_{\text{MR}\to\text{CT}}(G_{\text{CT}\to\text{MR}}(x)) - x||_1}{\exp\left(\log(\hat{\sigma_x})\right)} \right] + \frac{1}{2}\log(\hat{\sigma_x}) \right) + \left(\mathbb{E}_{y \sim I_{\text{MR}}} \left[\frac{||G_{\text{CT}\to\text{MR}}(G_{\text{MR}\to\text{CT}}(y)) - y||_1}{\exp\left(\log(\hat{\sigma_y})\right)} \right] + \frac{1}{2}\log(\hat{\sigma_y}) \right), \tag{3.10}$$

where, $\hat{\sigma_x}$ and $\hat{\sigma_y}$ are the predicted standard deviations of the CT volume and the MR volume. In practice, $\hat{\sigma}$ is obtained as an additional channel concatenated to the standard output predicted by the network. There are some notable changes in equation 3.10 compared to equation 3.9:

1. Instead of using the Gaussian prior for the loss function, the Laplacian prior is used because it results in a L_1 distance on the residuals (which is readily available within the cycle-consistency loss constraint). In practice, this means that instead of predicting the variance, the model now directly learns to predict the standard deviation.
2. The *logarithm* of the standard deviation is predicted.
3. The method of uncertainty estimation is identical for both forward (MR $\rightarrow$ CT) and backward (CT $\rightarrow$ MR) cycles.

The proposal is to make the model predict the logarithm of the standard deviation, in addition to the recovered volume that it already being computed for the original cycle-consistency loss. This, therefore, is called *aleatoric cycle-consistency loss*. As mentioned before, predicting aleatoric uncertainty attenuates the loss function, in that, the $\exp\left(\log(\hat{\sigma})\right)$ term in the denominator tempers the residual L_1 loss in the numerator. For inputs resulting in high uncertainty, this term reduces its direct effect on the loss. Using $\log(\hat{\sigma})$ rather than $\hat{\sigma}$ ensures that the model does not

predict high uncertainty for all inputs (meaning that it is ignoring the data), in which case, it is penalized as the contribution from the $\log(\hat{\sigma})$ term increases.

3.4.3 Unifying Epistemic and Aleatoric Uncertainties

This sections shows how the mathematical tools presented in sections 3.4.1 and 3.4.2 are combined to extract epistemic and aleatoric uncertainties from the existing 3D CycleGAN model.

Following the previously defined notation, let the generators in the forward and backward cycles be denoted by $G_{MR\rightarrow CT}$ and $G_{CT\rightarrow MR}$ and the input MR and CT volumes be I_{MR} and I_{CT}, respectively. Let $\hat{I}_{SynMR}$ and $\hat{I}_{SynCT}$ be the synthesized MR and CT volumes, and $\log(\hat{\sigma}_{SynMR})$ and $\log(\hat{\sigma}_{SynCT})$ be the predicted log standard deviations after translation. The model is trained with dropout using the same the adversarial and gradient consistency losses shown in equations 3.2, 3.3 and equation 3.7 respectively. Instead of using of the original cycle-consistency loss (equation 3.4), the updated aleatoric cycle-consistency (equation 3.10) is used, where, in addition to the recovered MR and CT volumes, their log standard deviations are also learned implicitly. During inference, the model weights are sampled from the approximate posterior $\hat{w} \sim q^*_\theta(W)$ to obtain the synthesized volumes along with the aleatoric uncertainties as follows[3]:

$$\left[\hat{I}_{SynCT}, \log(\hat{\sigma}_{SynCT})\right] = G^{\hat{w}}_{MR\rightarrow CT}(I_{MR}) \quad \text{and} \quad \left[\hat{I}_{SynMR}, \log(\hat{\sigma}_{SynMR})\right] = G^{\hat{w}}_{CT\rightarrow MR}(I_{CT}) \tag{3.11}$$

where, $G_{MR\rightarrow CT}$ and $G_{CT\rightarrow MR}$ are parameterized by the weights $\hat{w}$. Therefore, the output of a single generator provides both the synthetic volume and a measure of the aleatoric uncertainty.

For each test volume, T stochastic forward passes are performed through the network with dropout enabled (where, T refers to the number of MC samples). Each stochastic forward pass with weights $\{\hat{w}\}_{t=1}^{T}$ results in an unbiased estimate of the synthetic CT volume $\left\{\hat{I}^{\hat{w}_t}_{SynCT}\right\}_{t=1}^{T}$ and the aleatoric uncertainty map $\left\{\log\left(\hat{\sigma}^{\hat{w}_t}_{SynCT}\right)\right\}_{t=1}^{T}$. Then, the mean and variance of these T stochastic forward passes are computed, which are the predictive mean and model uncertainty (predictive

[3] It must be noted that only the forward cycle (left in the equation) is our primary concern during inference. The right equation is shown just for the sake of completeness.

variance), respectively. Mathematically, the predictive mean is $\mathbb{E}\left[\hat{I}^{\hat{w}_t}_{SynCT}\right] = \frac{1}{T}\sum_{t=1}^{T}\hat{I}^{\hat{w}_t}_{SynCT}$ and model uncertainty is $\mathrm{Var}\left(\hat{I}^{\hat{w}_t}_{SynCT}\right)$.

3.5 Experimental Protocol

This section lays out the practical details of the proposed methodology. The dataset is described in section 3.5.1, the training details are provided in section 3.5.2, and the method chosen for quantitative validation is presented in section 3.5.3.

3.5.1 The Dataset

This section gives specific details about the acquisition of the dataset and the preprocessing steps applied.

The MR and CT datasets were acquired from 3 different sources (2 for MR and 1 for CT). For MR, we used the dataset from the 2018 MICCAI Challenge on Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR (M3) Images [4]. This dataset consists of 16 volumes of the lumbar spine, comprising of 4 mutually aligned MR modalities for studying the effect of prolonged bed rest on lumbar intervertebral discs. The demographics of the subjects were not provided with the data. We chose to use only the water-phase images judging by a visual inspection of the contrast between the vertebral bodies and the surrounding regions. Our second source is a subset of the dataset described by Chevrefils *et al.* (2009) consisting of MRI 3-D multi-echo data volumes from 11 adolescent idiopathic scoliotic (AIS) patients with deformities ranging from mild to severe, acquired from CHU Sainte-Justine in Montréal, Québec. In particular, the dataset contained three patients' spine volumes with low severity (Cobb $12° - 24°$), four with moderate severity (Cobb $28° - 35°$), four with high severity (Cobb $43° - 60°$) deformations and focused more on the thoracic region (T1-T12) of the vertebral column (see figure 3.4 for reference). For CT, we used 2 sample volumes provided by 3D Slicer. In addition to the vertebral column, these volumes also contained the cardiac and the

[4] https://ivdm3seg.weebly.com/

abdomen regions, hence were cropped accordingly to focus only on the vertebral columns. Patient demographics were unavailable for the CT data also.

To ensure uniformity across the data, all MR and CT volumes were cropped and resized to $256 \times 128 \times 48$ sized volumes. The voxel sizes were resized to lie in $\{1, 1.5\} \times \{1\} \times \{1\}$ mm^3 for MRI data and $\{0.75, 1\} \times \{0.75, 1\} \times \{1, 1.5\}$ mm^3 for CT data. Out of 27 (16 MICCAI + 11 Scoliotic) original MR volumes, we had ground truth data for 5 scoliotic volumes obtained in the form of digitized landmarks reconstructed from preoperative biplanar X-ray images. Hence, these 5 volumes were used for testing. The rest of the volumes were used for training. Due to the lack of original MR and CT volumes, we used several data augmentation methods to increase the size of the training dataset. Initially, all the voxel intensities were normalized to lie in $[-1, 1]$. The data augmentation was entirely offline and included 3D rotation, Gaussian noise injection, elastic deformation, contrast stretching, and histogram equalization methods. The Gaussian noise was sampled from a normal distribution with zero mean and 0.01 standard deviation. In particular, a volume was taken and 10-15 additional volumes were obtained for each augmentation method. In total, we used 1112 MR volumes and 200 CT volumes for training including those synthesized by augmentation.

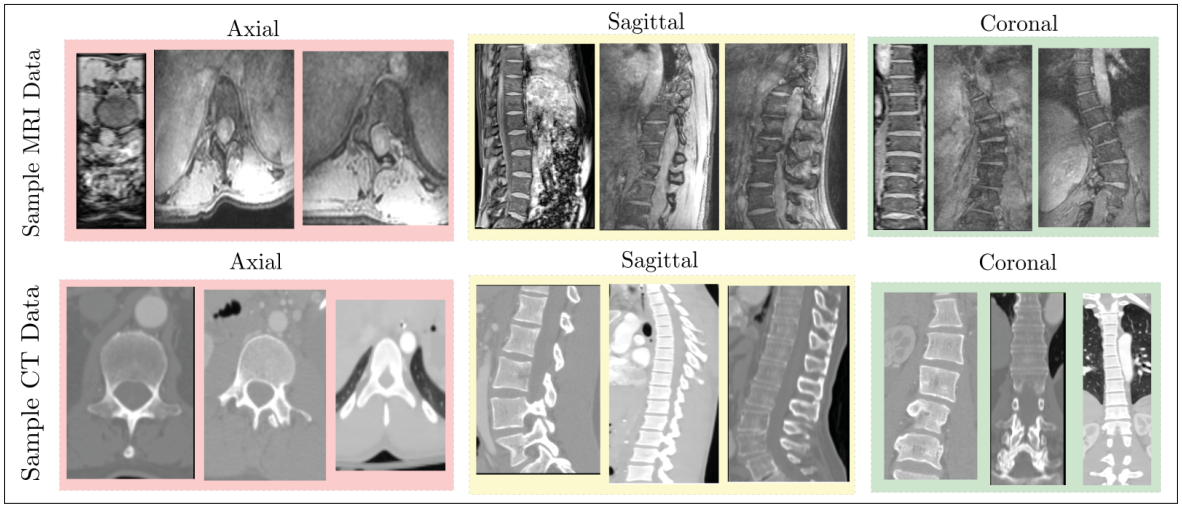


Figure 3.4 A few samples slices from the training data. *First row - axial, coronal and sagittal boxes*: first slice shows the MICCAI lumbar data, second and third slices show Scoliotic data with moderate (Cobb $28° - 35°$) and severe deformities(Cobb $43° - 60°$). *Second row*: CT data slices

Figure 3.4 shows a few instances of the training data. It is important to note the high degree of variability in the data, in terms of the severity of spinal curvature and different spine regions (lumbar and thoracic) that each MR dataset represents. This is in contrast to the CT data from 3D Slicer, which contains the entire vertebral column.

3.5.2 Training Details

We modified the U-Net (Ronneberger, Fischer & Brox, 2015) for the generator networks in our CycleGAN model to utilize the "U"-shaped architecture's specific downsampling, upsampling, and feature maps' concatenation properties. All 2D convolutional layers were replaced with 3D convolutional layers. For stable training, spectral normalization (Miyato et al., 2018) was used, followed by instance normalization (Ulyanov et al., 2016). We used ReLU and LeakyReLU activation layers in the downsampling and upsampling paths respectively. Inspired by the Zhang *et al.* (2018)'s architecture, we used 4 convolutional blocks, containing a sequence of 2 consecutive $3 \times 3 \times 3$ convolutions with stride 2 for each resolution such that maximum downsampling is 16 times for both the generators. For the upsampling path, instead of using transposed convolutions, which are known to produce checkerboard artifacts (Odena, Dumoulin & Olah, 2016), we used nearest-neighbor upsampling followed by a $3 \times 3 \times 3$ convolution. The PatchGAN model (Zhu *et al.*, 2017) was modified to be used as the discriminator. Each voxel in the final layer of the network had an overlapping receptive field of $46 \times 46 \times 46$ voxels, which was then classified as real or synthetic (instead of classifying the whole volume at once).

For optimization, the Adam solver (Kingma & Ba, 2014) was used with a batch size 2 and a learning rate of 0.0002. The model was trained from end-to-end for 200 epochs with linearly decaying the learning rate after the first 100 epochs. While the adversarial losses (section 3.2.1) are defined using the negative log-likelihood (NLL) objective, optimizing does not generally result in stable training. Therefore, as suggested by Zhu *et al.* (2017), the NLL objective was replaced with the least-squares loss (which is more stable and generates better results). Accordingly, the optimization objectives are changed as follows: the generator $G_{\mathrm{MR}\to\mathrm{CT}}$ now minimizes $\mathbb{E}_{y\sim I_{\mathrm{MR}}}\left[(D_{\mathrm{CT}}(G_{\mathrm{MR}\to\mathrm{CT}}(y)) - 1)^2\right]$, and the discriminator D_{CT} minimizes

$\mathbb{E}_{x\sim I_{\mathrm{CT}}}\left[(D_{\mathrm{CT}}(x)-1)^2\right]+\mathbb{E}_{y\sim I_{\mathrm{MR}}}\left[(D_{\mathrm{CT}}(G_{\mathrm{MR}\to\mathrm{CT}}(y)))^2\right]$. The training and inference were done on a single node of a remote compute cluster with 4 NVIDIA Tesla V100 32GB GPUs.

3.5.3 Quantitative Validation Method

In order to fully understand our approach for a quantitative evaluation, it is useful to understand the nature of the reference data. As mentioned in chapter 1, all scoliosis patients are required to undergo regular biplanar X-ray scan treatments to monitor the spinal curvature. These scans are utilized further for identifying important landmarks across vertebrae, which are then reconstructed in 3D. In this study, the 3D reconstructed landmarks obtained from the method proposed by Delorme et al. (2003) are considered to be the ground truths. An indirect method for evaluating the accuracy of the segmentation is by measuring the *point-to-surface* distance between the landmark points and the segmented surface using the iterative closest point algorithm. This part was done entirely on 3D Slicer using the Visualization Toolkit (VTK) graphical library.

The test set contained the data of five scoliotic patients, hence, the landmark points of those five patients were used for validation. Four of the five test patients (i.e. "Patient1" (P1), "Patient3" (P3), "Patient11" (P11), "Patient12" (P12)) had curvatures in the thoracic spine, whereas "Patient4" had a curvature extending to the lumbar spine. The number of landmarks also varied across the test data. In the cases of P1, P4, P11 and P12, all thoracic vertebral levels had 74 landmark points, except for vertebra T12, which had 72 landmarks. The lumbar vertebral levels (L1 and L2) had 66 landmark points each. In the case of P3, each thoracic level had only 17 landmarks (see figure 4.9, bottom-right). The patient data having 74 landmarks extensively covered the vertebra including the spinous process, transverse processes, lamina and pedicles of the vertebral arch, superior costal facets, superior and inferior articular facets, and the vertebral body. In contrast, patient data having 17 landmarks covered the spinous process, transverse processes, pedicles, and the vertebral body only.

It is important to note that the X-ray images were taken with the patient in standing position whereas the MRIs were acquired with the patient in prone position, causing significant differences

between the overall spine shapes (Harmouche *et al.*, 2012). However, the individual vertebrae within the spine remain rigid irrespective of the patient's posture. Hence, instead of validating on the entire vertebral column in one go, the point-to-surface distance for each vertebra was calculated separately. The ICP algorithm was used for initial registration of corresponding vertebrae before computing the mean distance error.

3.5.3.1 The ICP Algorithm

The iterative closest point algorithm (Besl & McKay (1992), Chen & Medioni (1992)) is mainly used to minimize the distance between two sets of point clouds (i.e. point-to-point) to align their shapes as closely as possible. Its extensions also include point-to-plane registration. Out of the two point clouds, one is said to be the source and the other is the reference. The reference data are generally fixed and the source is rigidly transformed using a series of rotation and translation operations to best match the shape of the reference. The goal of the ICP registration is to obtain a rigid transformation matrix such that the error defined by the root-mean-squared (RMS) distance between the source and the reference point clouds is minimal. While ICP helps in obtaining the best local alignment possible, it is heavily dependent on a good initialization. In cases where the data are arbitrarily initialized, there is a high chance that the algorithm may not converge to the optimal distance possible, irrespective of the number of iterations. The algorithm for ICP registration is as follows:

In this study, the ICP method was used to iteratively register the landmark points to the (individual) segmented surfaces of the vertebrae. The point cloud defined by the landmarks was considered to be the source and the vertebral surface was taken as the reference. Initially, the landmark points were loaded onto 3D Slicer as an array and their centroid was calculated. Then, the centroid of each segmented vertebra was computed by converting it into a voxel array obtained from the segment's binary labelmap image. This is followed by a translation operation to align the centroids of the landmarks and the segmented surface. However, it was found that the ICP algorithm could not find an optimal minimum by just aligning their respective centroids. Therefore, in order to obtain a good initialization for the ICP algorithm, the source point cloud

Algorithm 3.1 The ICP Algorithm

Input: *Landmark points (source) and segmented surface (reference), initial transformation of the source and the reference, stopping criteria N (in iterations).*
Output: *Rigid transformation aligning the points and the surface and mean distance (in mm).*
for *all iterations* $1 \rightarrow N$ **do**
    For each landmark point match the closest corresponding point in the segmented surface;
    Use translation and rotation operations to transform the set of points to align with the surface by minimizing the root-mean-squared distance;
    Realign the points by performing the rigid transformation;
end for

was roughly aligned to the reference surface once more by performing a rigid transformation using the "Transforms" module. Once after this alignment is done, the ICP algorithm was run for 100 iterations and the point-to-surface mean distance was computed. The results are shown in section 4.2. It should be noted this method does not require an equal set of points from the source and the reference data.

3.5.3.2 Point-to-Surface Distance

In general, studies discussing different methods for 3D reconstruction of the vertebrae of scoliotic subjects also extract relevant information from biplanar X-ray images in the form of a set of landmark points corresponding to anatomy of the vertebra to compute a metric called the point-to-surface distance (Mitulescu et al. (2001), Delorme *et al.* (2003)). It helps in better evaluating the global shape accuracy of the reconstructed vertebrae. As the name suggests, the goal of this metric is to simply calculate the distance between the properly aligned landmark points and the segmented surface.

The "Fiducials-to-Model Registration" module, an open-sourced extension of 3D Slicer, was used to compute the point-to-surface distance. As mentioned in the previous section, the ICP algorithm was run after aligning the landmarks and the segmented surface by using VTK's

"vtkIterativeClosestPointTransform" class. This aligned the source points to target surface in a least-squares sense. Once the source and the target were aligned, the point-to-surface distance was calculated automatically by the Fiducials-to-Model Registration module as follows:

1. A cell locator (vtkCellLocator) is initialized for efficiently locating 3D cells using an octree-based spatial search object.
2. For each landmark point, the "FindClosestPoint" method of the vtkCellLocator class is used to return the cell on the 3D segmented surface that is closest to the landmark point.
3. Then, the Euclidean distance is calculated between the landmark point and the surface cell.
4. Finally, the mean of the Euclidean distance considering all the landmarks is computed, which is the point-to-surface distance.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents the qualitative (sections 4.1 and 4.2) and quantitative (section 4.3) results obtained from our method. A general discussion is presented in section 4.4.

Several experiments were conducted to truly understand the effects of the gradient consistency (GC) loss and uncertainty estimations towards the quality of the translated CT volumes. The hyperparameter governing the importance of the GC loss (γ in equation 4.1) was also tuned to determine whether it especially improves translation at the boundaries. The outcomes of these experiments are shown further. It should be noted that the data of five scoliotic patients were used in the test set, hereafter referred to as "Patient1", "Patient3", "Patient4", "Patient11", and "Patient12". Their spinal deformities range from mild (Cobb $12° - 24°$) to severe (Cobb $43° - 60°$).

4.1 Qualitative Results - Translation

This section shows all the translation results obtained on the test data. Based on the quality of the translations and additional information obtained from the regions of uncertainty, the segmentations are obtained. Analyzing the translation results, especially the regions of high uncertainty, before Otsu segmentation is important because it gives the user an estimate of the amount of manual supervision required post segmentation.

4.1.1 Effect of varying γ

Let us recall the final optimization objective for training the 3D CycleGAN model:

$$\left(G^*_{\mathrm{CT}\to\mathrm{MR}}, G^*_{\mathrm{MR}\to\mathrm{CT}}\right) = \arg \min_{\substack{G_{\mathrm{MR}\to\mathrm{CT}} \\ G_{\mathrm{CT}\to\mathrm{MR}}}} \max_{\substack{D_{\mathrm{MR}} \\ D_{\mathrm{CT}}}} (\mathcal{L}_{\mathrm{CT}} + \mathcal{L}_{\mathrm{MR}} + \lambda \mathcal{L}_{\mathrm{Cycle}} + \gamma \mathcal{L}_{\mathrm{GC}}), \tag{4.1}$$

where λ and γ are the hyperparameters for weighting cycle- and gradient-consistency losses. Given the definition of the GC loss, it is natural to think that optimizing for this loss would help the model in translating the vertebral boundaries correctly. One can likewise hypothesize that increasing the value of γ in the above equation would only improve the translation accuracy. In order to visualize this empirically (and qualitatively), three models were trained independently with $\gamma = [0.5, 1.0, 5.0]$, where a different value fixed for each run. The uncertainties were also modeled in these three runs.

It was found that the model performance decreased with increasing γ values. In other words, the best translations were achieved when $\gamma = 0.5$. Figure 4.1 shows a comparison between a few slices of the translated volumes. Considering the red boxes shown on each slice, a qualitative analysis shows that certain regions of the spine, for instance, the spinous processes in Patient3, scoliotic curvature in Patient1, etc., have been better translated when $\gamma = 0.5$ compared to the other γ values. It can also be seen that $\gamma = 5.0$ results in relatively consistent bad translations, especially with the locations of the spinous processes. This can be interpreted as follows: successful translations from the CycleGAN model are heavily dependent on using the right combination of hyperparameters that lead to stable training. Since the primary objective of the model is to be cycle-consistent, increasing the influence of the parameters other than the cycle-consistency loss in the final objective heavily risks disturbing the equilibrium that the model achieves during the course of its training. As a result, it is unable to fully satisfy the cycle-consistency property, which in turn decreases the translation quality as γ increases. At this point, one might ask, what happens when the GC constraint is removed entirely? This is explored in the following subsections.

4.1.2 Experiments with the GC Loss and Uncertainty

Furthermore, four experiments were run to understand whether only the cycle-consistency constraint is strong enough for good translations (i.e. not including the GC loss) and whether making the model learn the uncertainty maps on its own improves its performance meaningfully.

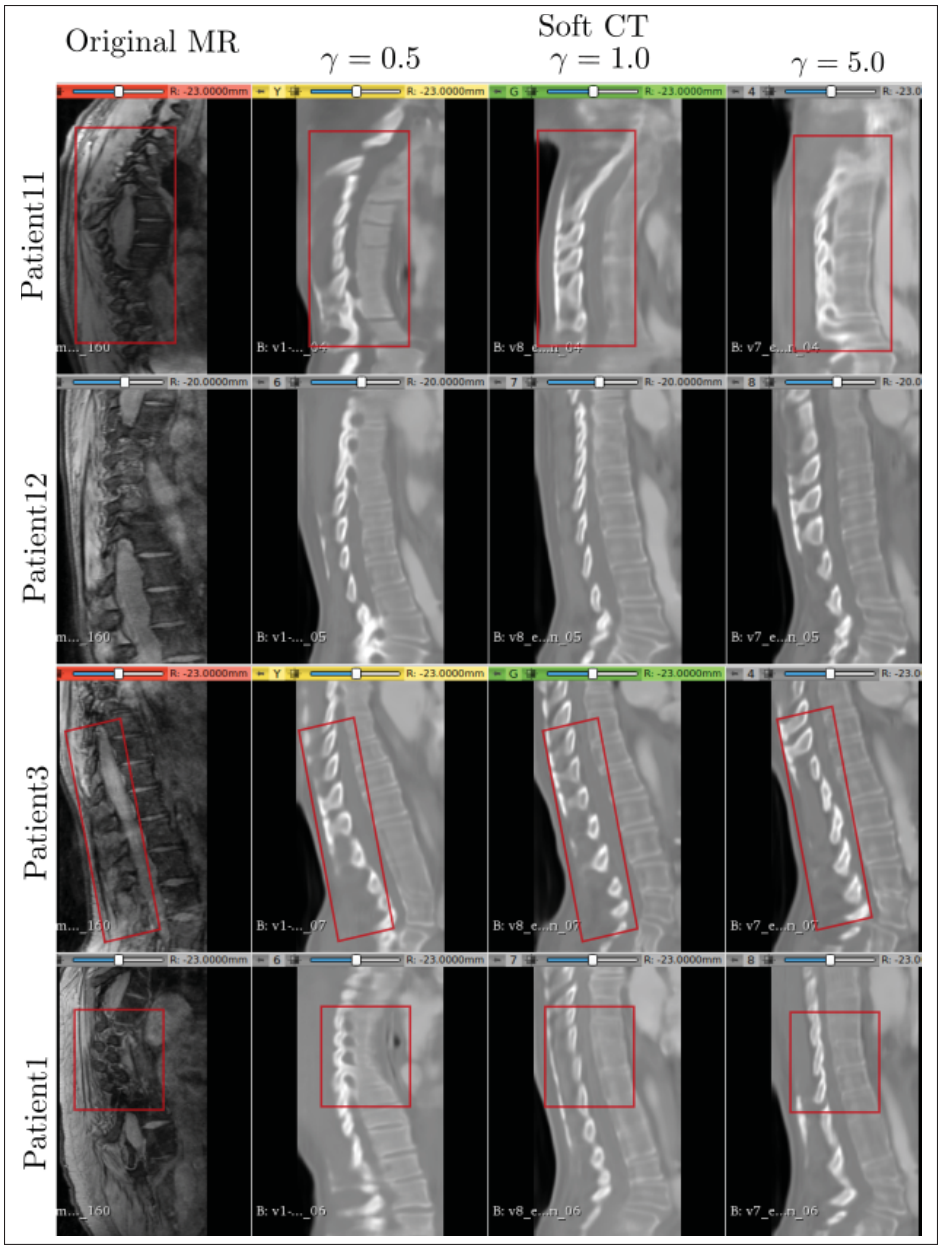


Figure 4.1 The effect of varying γ in translations. Left-to-right: original sagittal MR slices, "soft" CT predictions with varying γ values. Red boxes compare specific regions across the translation results from different γ values. Top-to-bottom: Results for Patient11, Patient12, Patient3, and Patient1

To that end, all four variations of the above two combinations were modeled independently and named as follows:

1. **WithGC_withUnc** - where, a model is trained with both GC loss and uncertainty computations enabled,
2. **WithoutGC_withUnc** - where, a model is trained without the GC loss but provides uncertainty estimations,
3. **WithoutGC_withoutUnc** - where, a model is trained *without* both GC loss and uncertainty computations (i.e. the default CycleGAN), and
4. **WithGC_withoutUnc** - where, a model is trained with GC loss but without the uncertainties.

In the models where the uncertainties are estimated, 20 MC samples were used during inference to obtain the final predictions. Let us now compare different MR-CT translations between the above combinations qualitatively in the following subsections. However, before diving into the results, it is important to understand the distinction between "soft" and "hard" predictions (translations). Recall that by the virtue of estimating the epistemic uncertainty, a fixed number of MC samples (20, in this case) are taken from the approximate posterior distribution. Therefore, the final translation is the mean of these 20 MC samples. As a result, the inconsistencies across the MC samples get smoothed out and a "soft" prediction is obtained. This is in stark contrast to the conventional inference strategy, where only the best set of weights is used for translation, resulting in a "hard" prediction.

4.1.2.1 Effect of the GC Loss without Uncertainty

This subsection compares the results of the models (3.) withoutGC_withoutUnc and (4.) withGC_withoutUnc. Figure 4.2 shows the ("hard") translations obtained for Patient1, Patient3, and Patient12 from the test set.

Considering the red and green arrows in the 1st and 3rd rows of figure 4.2, it is clear that optimizing for gradient consistency during training helps the model learn the vertebral shapes. In other words, it helps localize the bone structures from the training volumes. Also, the absence of uncertainty estimates means that the model is confident in all its "hard" predictions throughout, however, that is not the case where the translations are incorrect. This suggests that model

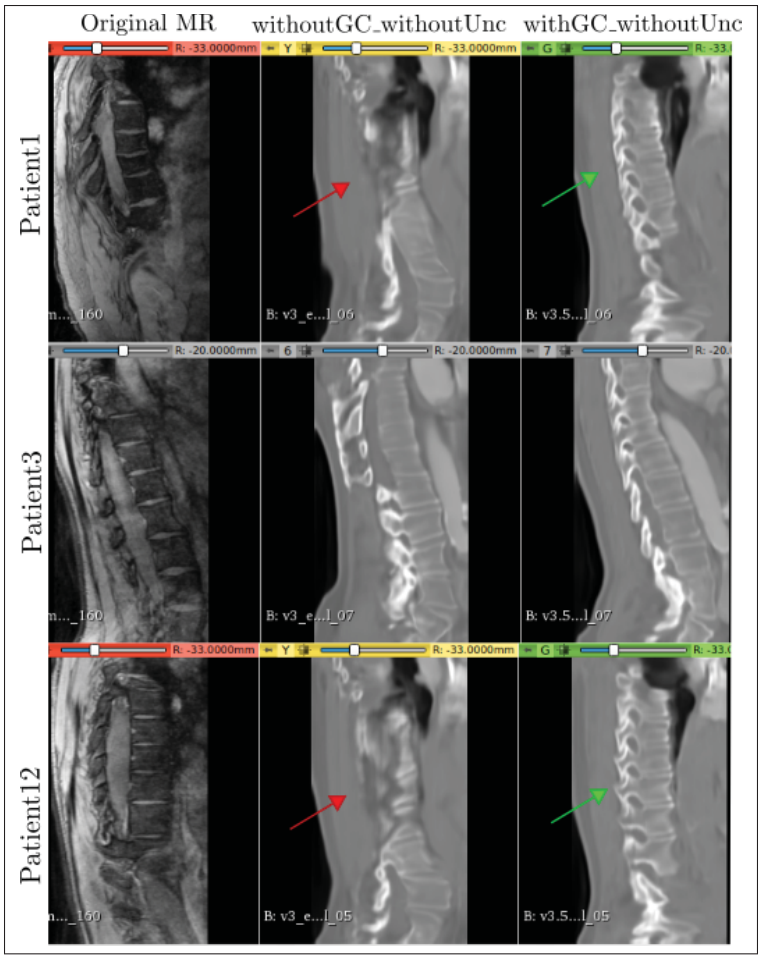


Figure 4.2 "Hard" Synthesized CT predictions for Patient1, Patient3, and Patient12. Left-to-right: Original MR slice, synthesized CTs without uncertainty for models without and with GC, respectively. Green and red arrows show the difference in translation with and without GC, respectively. "Unc."= Uncertainty

uncertainty becomes extremely important in post-processing tasks (such as segmentation, in this study).

4.1.2.2 Effect of the GC Loss with Uncertainty

In this subsection, the results of models (1.) withGC_withUnc and (2.) withoutGC_withUnc are compared. Figure 4.3 shows the translations and the uncertainty estimates for Patient1 and

Patient3 from the test set. Recall that "soft" predictions differ from "hard" predictions in that they obtain a smoothed version of all the MC samples.

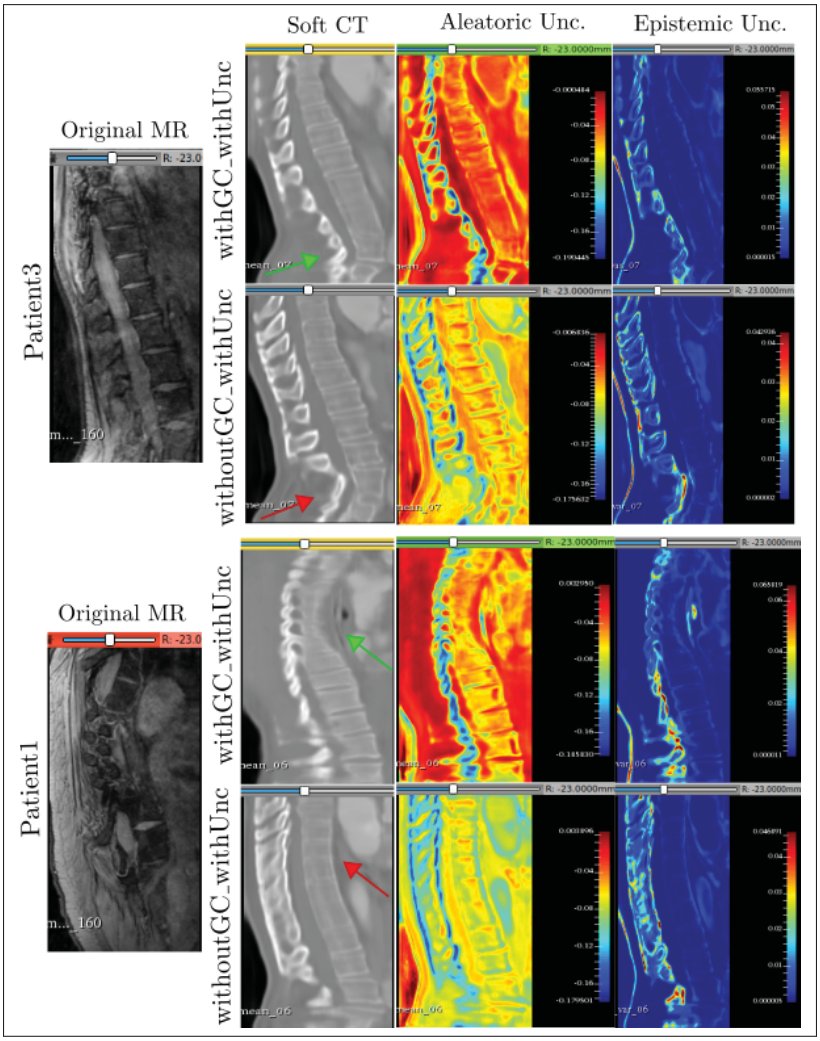


Figure 4.3 The soft translations along with aleatoric and epistemic uncertainties. Left-to-right: original sagittal MR slices for Patient1 and Patient3, "soft" CT predictions, aleatoric maps learned by the models, and epistemic uncertainties. Top-to-bottom: 1st and 3rd rows show results with GC, 2nd and 4th rows show results without GC. Green arrow points to the spinal curvature better translated with GC and red arrow shows the same region translated without GC. Blue and red regions in the uncertainty maps refer to low and high uncertainties respectively. "Unc."= Uncertainty

Considering the bottom-half of figure 4.3, it can be seen that spinal curvature of Patient1 has been slightly better captured by the model that was trained with the GC loss (shown by green

and red arrows). For Patient3 (top-half of figure 4.3), the bottom thoracic vertebrae have been better translated with GC (row 1) compared to the model trained without the GC loss (row 2). There are three observations regarding the uncertainty maps:

1. The epistemic and aleatoric uncertainties seem to complement each other in that, the former shows high uncertainties near the spinous process boundaries, while the latter covers all regions but the spinous processes.
2. Recall from equation 3.10 that the aleatoric maps were learned by comparing the recovered MR volumes with the original MR volumes. Also, in order to satisfy cycle-consistency, the recovered MR volumes are solely based on the quality of the synthetic CT volumes from the forward cycle where the soft-tissue information is lost and bone structures are in focus. Therefore, the high uncertainty corresponds to the soft tissue regions lost during the forward cycle translation going from MR to CT. Notice that the soft tissue regions in rows 1 and 3 (withGC) are fully red (highly uncertain), whereas the bones are in yellow and blue (relatively less uncertain). On the other hand, since the epistemic uncertainty depends only on the model parameters, it specifically shows that the model's confidence is low in translating the spinous processes.
3. In the case without GC (rows 2 and 4), it appears as if the model was unable to distinguish between the bones and the soft tissues, hence predicting similar aleatoric uncertainty (yellow/green regions) across the entire image.

It must be noted that by the virtue of modelling the aleatoric uncertainty, the model learns to distinguish between the soft tissue and bone regions by itself without any external conditioning.

4.1.2.3 Effect of Modeling Uncertainty with GC Loss

This subsection compares the results of the models (4.) withGC_withoutUnc and (1.) withGC_-withUnc. Figure 4.4 shows the "hard" translations and the "soft" translations along with uncertainty estimations for Patient4, Patient12, and Patient11 from the test set. Before analyzing the results, it is important to appreciate the amount of additional information the uncertainty estimates provide, compared to their non-uncertainty counterparts. Consider the green boxes

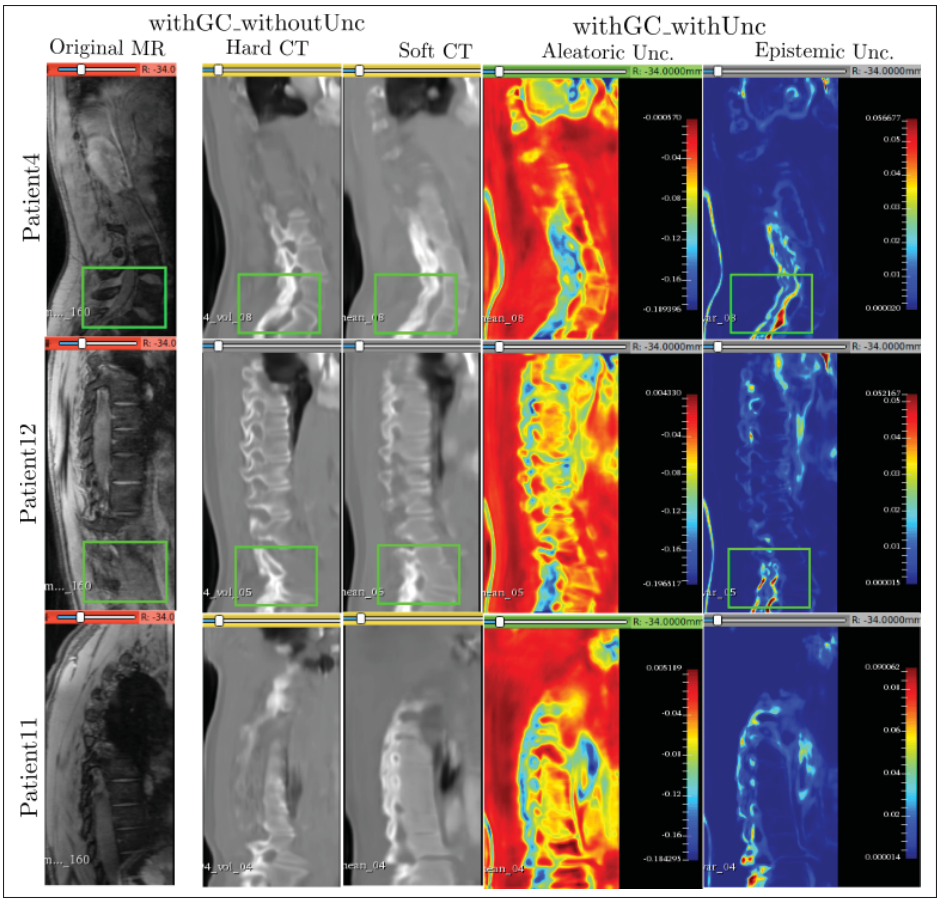


Figure 4.4 The "hard" and "soft" CT translations along with aleatoric and epistemic uncertainties for the latter. Left-to-right: Original MR slices, "hard" CT results for the model without uncertainty, columns 2, 3, and 4 - mean prediction with aleatoric and epistemic maps, respectively. Green boxes show the specific regions compared across translations. Blue and red regions in the uncertainty maps refer to low and high uncertainties respectively. "Unc" - Uncertainty

across all slices in Patient4 and Patient12: it can be seen that the translation from the MR slice has not been accurate to translate the spinous processes. Both "hard" and "soft" predictions are similar to each other, however, the model trained with both GC and uncertainty conveys that it is not confident about its translation of the spinous processes. This can be seen in the form of high epistemic uncertainty within the green boxes. Therefore, the user is alerted to be careful around this region during the post-processing of the segmented vertebral body.

4.1.2.4 Effect of Modeling Uncertainty without GC Loss

This final subsection compares the results of the models (3.) withoutGC_withoutUnc and (2.) withoutGC_withUnc. Figure 4.5 shows the corresponding results for Patient3, Patient4 and Patient12.

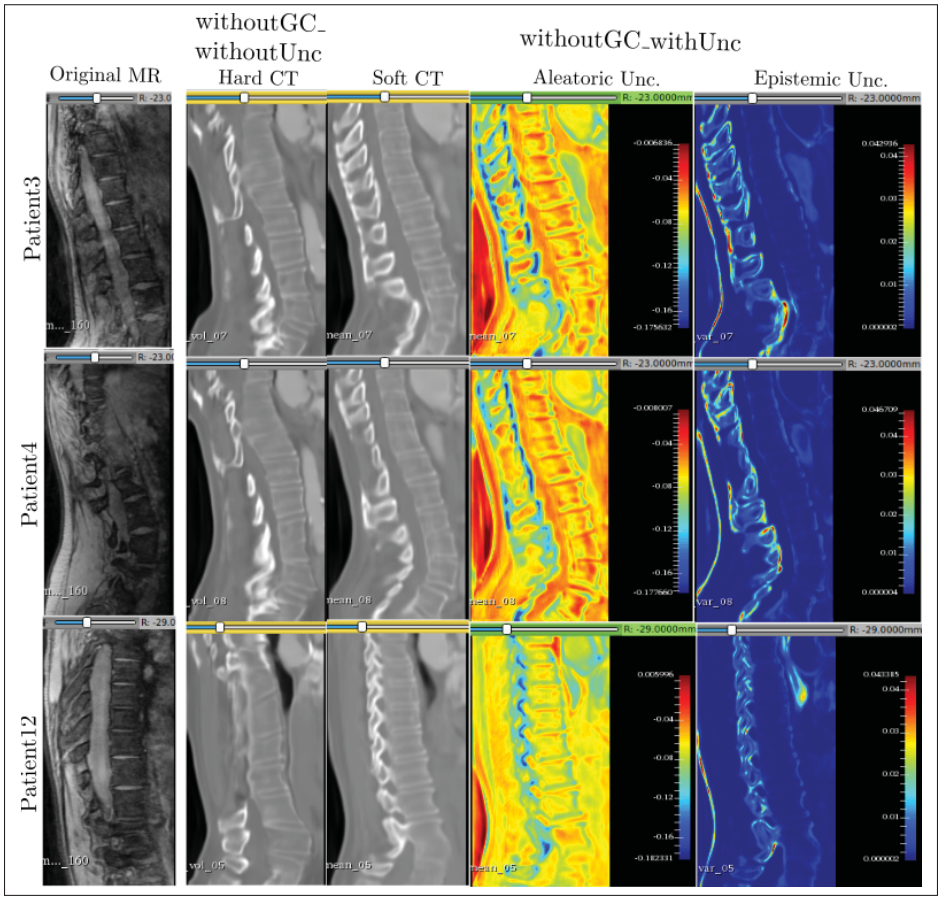


Figure 4.5 The "hard" and "soft" CT translations along with aleatoric and epistemic uncertainties for the latter. Left-to-right: Original MR slices, results for the model without uncertainty, columns 2, 3, and 4 - mean prediction with aleatoric and epistemic maps, respectively. Blue and red regions in the uncertainty maps refer to low and high uncertainties respectively. "Unc" - Uncertainty

Considering the first two rows of figure 4.5 containing the results of Patient3 and Patient4, it can be seen that the "hard" CT translations appear to be similar for the model trained without both GC and uncertainty. However, due to the absence of uncertainty information in this case, it

is difficult to understand where the model might have translated incorrectly. While the "soft" translations themselves are not perfect (Patient12 in fig. 4.5), they are able to better capture the shapes of the spinous processes, relatively (Patient3 and Patient4 fig. 4.5). In addition, considering the "hard" CT slices of Patient3 and Patient4, it seems as if depriving the model of GC and uncertainty constraints has affected its ability to learn the vertebral structure specific to the patient and output generic translations unlike its "soft" CT counterpart. Therefore, it can be said that by modelling aleatoric uncertainty during training, the model tries to offset the lack of GC. However, bone localization still remains an issue. Also, the epistemic uncertainty focuses on the spinous processes, thereby making it easier to rectify the incorrect translations during manual post-processing.

4.1.2.5 Takeaways

In the four experiments described above, we have seen how the GC loss and uncertainty estimations play a role during and after training in the quality of the synthesized CT volumes. Some of the important observations are noted in this subsection that would help during Otsu segmentation, the results of which are shown in the following section.

Out of all the four experiments, the results obtained by the model "withoutGC_withoutUnc" have been the least positive (figure 4.2). This suggests that the default CycleGAN model, without GC and uncertainties is not enough for translating bone intensities across MR and CT volumes satisfactorily. Moreover, modeling none of the uncertainty measures affected the model's performance substantially. The effect of modeling just the uncertainties without optimizing for the GC loss can be clearly observed from the results of the "withoutGC_withUnc" experiment (figure 4.5). The vertebral shapes in general, were better captured compared to its "withoutUnc" counterpart. This seems to be the case mainly because modeling aleatoric uncertainty acts as learned loss attenuation (Kendall & Gal, 2017), thereby preventing the model from predicting those recovered MR volumes that are very far from the original MR volumes during training (see equation 3.10). Also, recall from section 4.1.2.1 (figure 4.2) that GC helps in localizing the bone structures during training. Therefore, this leads to our first important observation that

having either the GC constraint or the uncertainty estimations is paramount for the quality of the translations.

On a few test volumes, it was also observed that having the model optimize for GC helps it better capture the spinal curvature (figure 4.3). The aleatoric uncertainty maps between the experiments "withGC_withUnc" and "withoutGC_withUnc" also showed different levels of uncertainty. In the former, optimizing for GC helped the model localize the vertebral structures, which in turn helped the model to learn to automatically distinguish between the bones and the soft tissue regions (as shown with yellow and red regions of aleatoric uncertainty respectively). In the latter, the lack of distinction between the bones and soft tissues (due to the lack of GC) resulted in nearly uniform aleatoric uncertainty. The corresponding epistemic uncertainty results specifically show high uncertainty in the spinous processes, thereby making it a target requiring increased supervision during post-processing. This leads to two more important observations: (i) despite minor differences in the translations, it is better to optimize for the GC loss, in addition to modeling the uncertainty estimates, as long the training remains stable and memory constraints allow, and (ii) out of the two uncertainty maps, providing the user only with epistemic uncertainty is more useful for post-processing tasks, while the aleatoric uncertainty helps the model identify and distinguish different regions in the training data, leading to improved performance.

One can also observe that the regions of high epistemic uncertainty have mostly been near the boundaries of the spinous processes. Since epistemic uncertainty is concerned only with the models' parameters (and not the input data), this can be understood as the models' inability to correctly process the locations of the spinous processes from the input MR volumes, thereby making them hard to detect and eventually leading to uncertainty. This seems natural because the spinous processes do not typically lie on the same horizontal levels as the vertebrae and their locations cannot be easily identified in the case of scoliotic spines with moderate to severe curvatures. It is also useful to note that the existing literature does not talk about the segmentation of the spinous processes in MR images. Therefore, this again reinforces the fact that uncertainty estimations do help in extracting more information out of the unsupervised CycleGAN model.

4.2 Qualitative Results - Segmentation

Following the observations from the previous section, the translation results obtained from the "withGC_withUnc" experiment were used for Otsu segmentation and quantitative validation. Figures 4.6 and 4.7 show the Otsu segmentation and the 3D models for all the five test patients. Taking into consideration the high degree of variability in the training dataset in terms of the presence of thoracic and lumbar vertebral levels, the varying spinal curvatures of scoliotic patients, and most importantly the lack of the corresponding ground truths in the translated CT domain, the model obtained decent translation results. Despite having only 6 original scoliotic volumes for training, the spinal curvature was well-captured by the model. The Otsu segmentation in itself was found to be insufficient for reconstructing a good 3D model of the scoliotic spine, in that it typically resulted in over-segmentation, including the intervertebral disks and the surrounding non-spine anatomy. Therefore, manual post-processing was done for all the segmented test volumes. It was entirely carried out on 3D Slicer using the paint-brush tool in the "Segment Editor" (default) extension. In particular, several over-segmented regions were cut out and median smoothing using a 3×3 kernel was also performed.

Some interesting observations were made during the post-processing. It was found that the model consistently mis-segmented the spinal cord to be the bones, mainly the transverse processes and laminae regions. As a result, the exact locations of the spinous processes seemed to be shifted by the width of the spinal cord. Recall from section 4.1.2.5 that the regions of high epistemic uncertainty were concentrated around the spinous process boundaries. Therefore, this suggests that the model was inherently unsure about the existence and the location of the spinal cord, hence high epistemic uncertainty in those regions. Due to excessive curvature in certain regions of the spine, the model was unable to translate the intensities accurately (shown by the red boxes in figures 4.6 and 4.7). Such regions were directly assumed to be the bones resulting in over-segmentation, which had to be segmented manually using additional supervision from the input MR volumes. While any kind of manual intervention is undesirable, it must also be noted that even in the case of the real CT volumes, manual intervention was necessary to focus only the vertebral column and remove the over-segmented non-spine anatomy. Figure 4.8 shows an

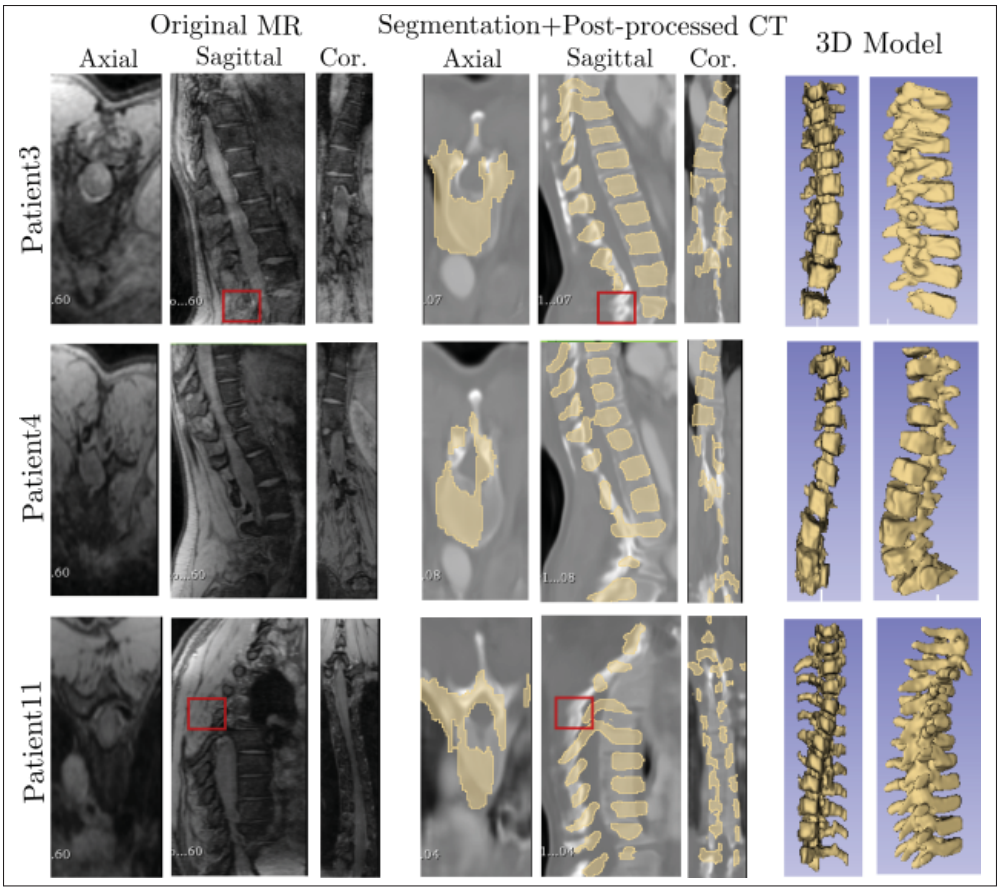


Figure 4.6 Segmentation and 3D spine models for 3 test patients. Left-to-Right: Axial, Sagittal and Coronal slices of the original MR and post-processed CT respectively. Final column shows the 3D model of the spine. Top-to-Bottom: Corresponding results for Patient3, Patient4, and Patient11. Red boxes show the mis-translated and mis-segmented regions corrected after manual post processing. "Cor." - Coronal

example of an Otsu-segmented CT volume before and after the manual correction. Though with varying amounts of supervision, since both real and synthesized CT volumes require manual intervention, this suggests that our primary objective of devising a cross-modality synthesis method for segmenting MR volumes as easily as their CT counterparts, is achieved. Moreover, modelling uncertainty has also led to increased model performance and user-interpretability. In particular, the regions of high-epistemic uncertainty provided targets requiring special attention, hence improving the overall segmentation. While aleatoric uncertainty maps are not as easily interpretable to the user, making the model learn to differentiate between the anatomical regions

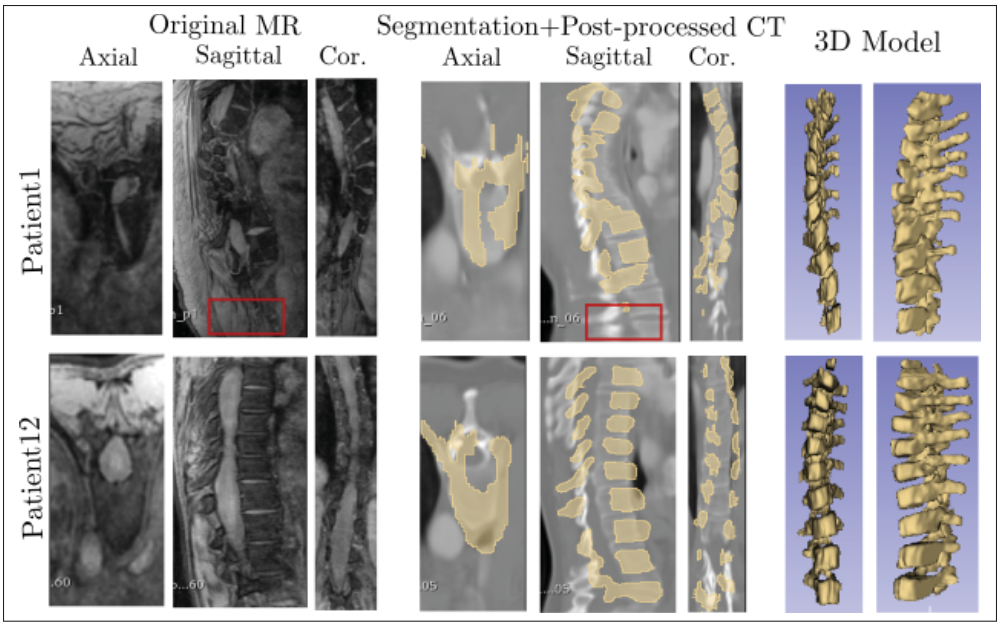


Figure 4.7 Segmentation and 3D spine models for the remaining 2 test patients. Left-to-Right: Axial, Sagittal and Coronal slices of the original MR and post-processed CT respectively. Final column shows the 3D model of the spine. Top-to-Bottom: Corresponding results for Patient1 and Patient12. Red boxes show the mis-translated and mis-segmented regions corrected after manual post processing. "Cor." - Coronal

even when the soft tissue information is lost during the forward translation, is useful. Therefore, this achieves our secondary objective of increasing model interpretability and segmentation performance as a consequence of estimating uncertainty.

4.3 Quantitative Results

In order to evaluate the accuracy of the 3D reconstructed vertebrae, the anatomical landmark points from the corresponding biplanar X-ray images for the five patients in the test set and the surfaces of the segmented vertebrae were compared by calculating the point-to-surface distance (section 3.5.3.2). Figure 4.9 shows an example of the anatomical landmarks overlayed on the segmented vertebra before and after performing the registration. The "Before Registration" stage in figure 4.9 shows the result after aligning just the centroids of the landmark points and the segmented surface. As mentioned in section 3.5.3.1, directly computing the point-to-surface

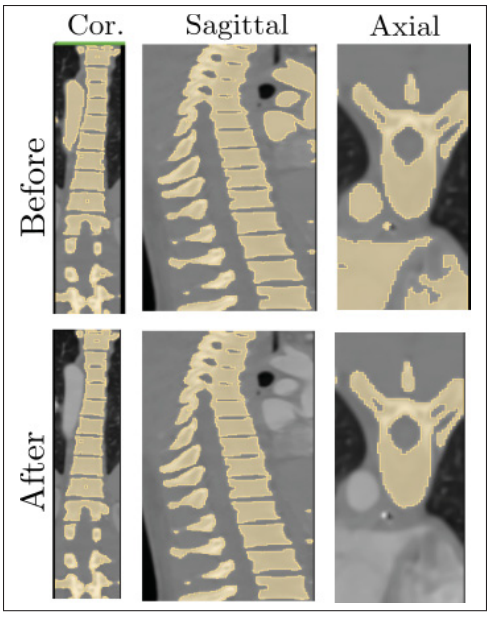


Figure 4.8 Necessity of manual intervention in a real CT volume. Top row shows the result of an Otsu segmentation, the bottom row shows the results after cutting out the non-spine parts."Cor." - Coronal

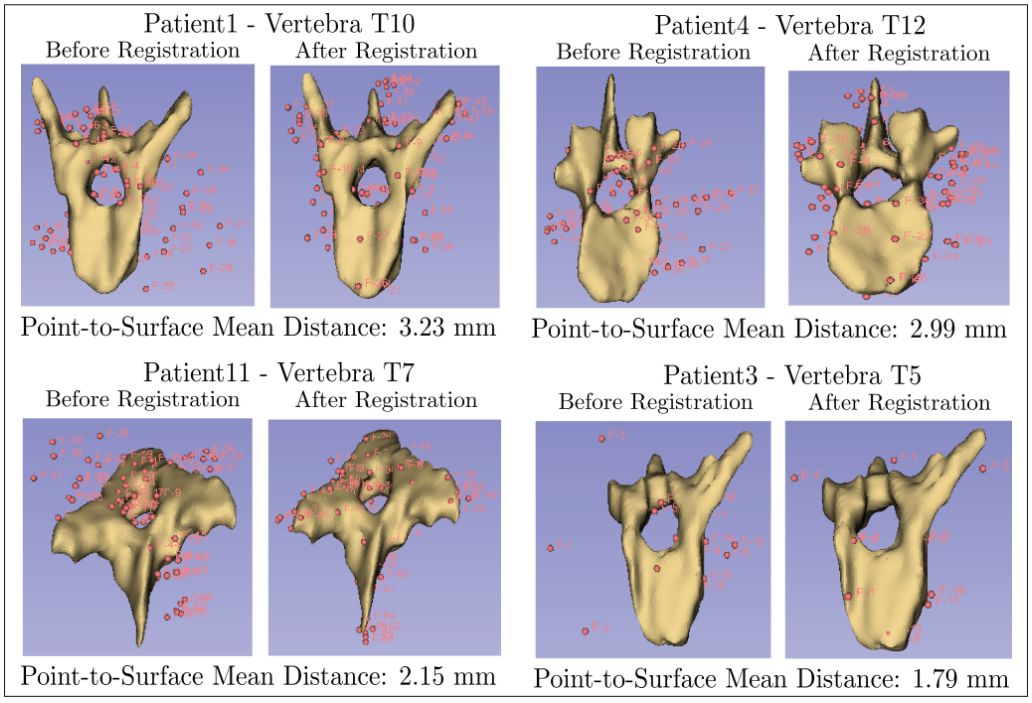


Figure 4.9 Anatomical landmarks (in pink) overlayed on the individually segmented vertebra. Notice that before registration, mainly the spinous and transverse processes, and the vertebral bodies are not aligned. A rigid transformation, followed by ICP registration, aligns the two shapes and the mean distance between them is computed

distances from this stage resulted in higher errors. Therefore, from the "Before Registration" stage, the landmarks were roughly aligned once more by performing a few translation and rotation operations (rigid transformation). The ICP algorithm was then run for 100 iterations and the point-to-surface mean distance was computed. Table 4.1 shows the resulting distance errors for different vertebral levels.

Table 4.1 Accuracy of the 3D reconstructions of the vertebrae shown for 46 scoliotic vertebrae in 5 patients (lower the distance the better)

Level	**Point-to-surface mean distance (mm)**					**Mean per Level (mm)**
	Patient1	**Patient3**	**Patient11**	**Patient12**	**Patient4**	
T2	-	-	5.42	-	-	5.42
T3	5.15	-	3.17	2.60	-	3.64
T4	4.02	2.31	2.72	2.96	-	3.00
T5	4.23	1.79	3.65	3.43	-	3.28
T6	4.64	1.88	2.49	3.09	-	3.02
T7	4.28	1.96	2.15	3.0	2.25	2.73
T8	5.05	1.96	2.29	3.37	2.65	3.06
T9	3.37	1.37	2.89	2.76	3.82	2.84
T10	3.23	1.40	2.31	2.69	3.67	2.66
T11	4.33	1.56	3.10	3.63	3.29	3.18
T12	5.66	-	3.63	-	2.99	4.09
L1	-	-	-	-	3.42	3.42
L2	-	-	-	-	4.20	4.20
Mean ± S.D	4.40 ± 0.72	1.78 ± 0.3	3.07 ± 0.89	3.06 ± 0.34	3.29 ± 0.6	
Total	**46 Vertebrae** with a mean error of **3.17 ± 1.04 mm**					
	(w/o P3) **38 Vertebrae** with a mean error of **3.46 ± 0.89 mm**					
	(w/o P3 & P4) **30 Vertebrae** with a mean error of **3.51 ± 0.94 mm**					

There are two important points to be noted considering table 4.1: (i) Not all test patients had curvatures in the same region of the spine which explains the absence of point-to-surface distance for certain vertebral levels, and (ii) One can also observe that the distance errors at the initial and the final vertebrae are consistently higher than the those in the middle. This is because these vertebrae were abruptly cut-off from the MRI images. A common observation is that all the patients have at least some degree of curvature between vertebral levels T7-T11. A possible alternative for improving the accuracies of the initial and final vertebral levels is to include a set of buffer vertebrae both above and below the region of interest such that its translation is improved.

It can be seen that the mean distance error for Patient1 is the highest with 4.40 ± 0.72 mm. Also, Patient1 suffers from a severe spinal deformity (Cobb $43° - 60°$). This confirms that the model struggles to translate the bone intensities in scoliotic MR volumes with severe spinal curvature. Such high anatomical variation resulted in a poor synthesis of the corresponding CT volume, hence affecting the subsequent Otsu thresholding and 3D reconstruction.

On the other hand, the mean distance error for Patient3 is the least with 1.78 ± 0.3 mm. This is primarily due to the fact that only Patient3 had significantly fewer landmark points (17) compared to the rest of the patients' data (~70). As a result, the lesser the number of landmarks the lower is the noise in terms of matching the landmarks' locations to the nearest surface, hence resulting in smaller point-to-surface distances. This is illustrated in the bottom-right panel of figure 4.9. Therefore, for a fair comparison, the second-last row of table 4.1 shows the total mean point-to-surface distance both with and without Patient3's scoliotic vertebrae. It should also be noted that out of the five test patients, the distance error magnitudes are similar for Patient4, Patient11, and Patient12 averaging around 3 mm.

The total point-to-surface distances shown in the table are comparable (with less variance, however) to Delorme *et al.* (2003) who used Direct Linear Transformation (DLT) and geometrical kriging to reconstruct vertebrae from landmark points digitized from 2D planar radiographic images and reported a global error of 3.3 ± 3.8 mm for 60 scoliotic vertebrae. It must be noted that the accuracy of their method was evaluated by comparing the reconstructed vertebrae with that obtained from scoliotic CT scan images, which are widely considered to be the gold standard due to their superior resolution. Since our method solely uses scoliotic MR volumes (with poorer resolution compared to CT volumes) for reconstructing the vertebrae, it is natural that our distance error is relatively higher in the case of total mean point-to-surface distance without Patient3's results. The fact that the average point-to-surface distance error is on a similar order of magnitude to 3D landmark reconstruction error suggest that at least some of our error inherently accounts for a possible error in the digitization of these ground truth landmarks, and therefore is not solely based on our method of vertebrae reconstruction.

In our previous work (Naga Karthik *et al.*, 2021), the total distance error of 3.41 ± 1.6 mm for 28 scoliotic vertebrae among three scoliotic patients (Patient1, Patient11, and Patient12) was reported. Comparing it with the proposed method, it is observed that two more vertebrae could be segmented for the same patient cohort and a combined mean point-to-surface distance of 3.51 ± 0.94 mm for 30 scoliotic vertebrae is obtained. While the mean distance error is slightly higher, it is offset by lower variance across all the vertebrae. We believe that low variance in the results is a direct consequence of modelling uncertainties, which provide useful diagnostic information to the user and the model itself. Moreover, by just comparing the result for Patient1 having the highest point-to-surface distance due to a severe curvature, it is observed that the current method obtains an error of 4.40 ± 0.72 mm for 10 vertebrae over our previous work reporting an error of 4.45 ± 0.93 mm for 9 vertebrae. This conveys that the proposed method is able to better capture the spinal curvature in severe scoliosis cases and also segment more vertebrae.

4.4 Discussion

While the proposed method obtained better results compared to our previous work in terms of segmenting more vertebrae and providing uncertainty estimates, there exist a few limitations:

- It must be noted that the segmentation stage is not automatic and requires manual post-processing after performing the Otsu segmentation. The lack of ground truth scoliotic CT volumes makes it difficult for the synthesized CT volumes to be automatically segmented. However, to eliminate the post-processing step, one can use semi-automatic thresholding methods that take user-initialized seed points on the synthesized CT volumes as the input and segment vertebral bones accordingly.
- Out of the two uncertainties modeled, only the epistemic uncertainty can be meaningfully interpreted by the end user. This is because the aleatoric uncertainty is typically obtained by comparing the model prediction with the actual ground truth, which, under ideal circumstances, would mean that the synthesized CT volume is compared with the ground truth CT volume. However, in order to circumvent the issue of the lack of ground truth data,

the proposed method compared the recovered MR volume with the original MR volume. Since the soft tissue information is already lost in the forward translation, the aleatoric uncertainty map focuses more on identifying different tissues surrounding the spine, hence is not easily interpreted by the user.

- Another issue common across all the synthesized CT volumes is the consistent mis-segmentation of the spinal cord as the vertebral bones, thereby suggesting the model's inability to locate the spinal cord from the training data.

CONCLUSIONS AND FUTURE WORK

A brief summary of the important contributions is given in section 5.1 and some directions for future work are listed in section 5.2.

5.1 Contributions

This thesis presented a novel method for the 3D segmentation of the scoliotic spine using cross-modality synthesis and also proposed computationally-efficient technique for modelling aleatoric uncertainty in the unsupervised learning setting. The major contributions can be summarized as follows:

- A 3D CycleGAN model was trained on unpaired MR-CT data instances for a volume-to-volume translation across the MR-CT domains. This approach leveraged the fact that the bone intensities in CT data are highly detailed compared to its MR counterpart, which are more suitable for studying soft-tissues surrounding the bones. The standard CycleGAN model was augmented by optimizing for gradient consistency resulting in well-localized vertebral boundaries. The Otsu's method, a simple, intensity-based thresholding algorithm was then applied to the translated CT volumes for segmentation. The over-segmented regions arising from the translation errors caused by the lack of sufficient data and unsupervised training, were manually corrected during post-processing. A 3D model of the spine was finally generated using the post-processed synthetic CT volume. The point-to-surface distance was used as quantitative validation metric. A preliminary part of this work was presented at the SPIE Medical Imaging Virtual Conference, 2021, as "Three-dimensional Segmentation of the Scoliotic Spine from MRI using Unsupervised Volume-based MR-CT Synthesis" (Naga Karthik *et al.*, 2021), where the data of three scoliotic patients was used and a mean distance error of 3.41 ± 1.06 mm for 28 individual vertebrae was reported.
- The unsupervised nature of the problem arising due to the lack of ground truth scoliotic CT data severely limits our ability to understand the model's behaviour and the confidence in its

predictions. Hence, this work was immediately followed upon and a Bayesian adaption of the 3D CycleGAN model was proposed in order to extract the uncertainty information in terms of aleatoric and epistemic uncertainties. In particular, a novel adaption to the cycle-consistency loss, termed as *aleatoric cycle-consistency loss* was proposed, which helped in obtaining both the translated CT volumes and aleatoric uncertainty maps using the pre-existing generator network. Several experiments were performed to determine the effectiveness of optimizing for gradient consistency and modelling uncertainties. A total of five scoliotic patients' data were used for validation. The point-to-surface mean distance was found to be 3.17 ± 1.04 mm for 46 scoliotic vertebrae.

In terms of clinical functionality, the proposed methodology allows us to obtain 3D models of scoliotic spines using a non-invasive imaging modality that could help in fusing the intraoperative images and guide the surgeon during minimally invasive interventions. This will eventually make such interventions safer and also improve the quality of the patient's life.

Also, when the proposed method is validated across a larger cohort of patients, the generated 3D models will allow the registration of the MR volumes with the 3D spine models reconstructed from the biplanar X-ray images. The advantage of such a multi-modal fusion is that a complete 3D model of the patient's trunk along with the soft-tissue information and the underlying bone structures can be generated. This will be useful for surgery planning as the surgeons will have the option of choosing the best treatment strategy based on the post-operative appearance of the trunk and can also take the patient satisfaction into account.

5.2 Future Work

The depth and the difficulty of the problem provide some useful directions for future work. As mentioned before, a recurring problem across the translations was the mis-segmentation of the spinal cord. In order to prevent such issues, the segmentation of the spinal cord could be done

initially, and the segmented volumes could be fed as the input to the model so that it does not get confused during the actual MR-CT translation. This could also lead to lower uncertainties near the spinous process regions.

The proposed method also struggled in the case of scoliotic spines with excessive curvature requiring more manual intervention. In such cases, one can preemptively segment various structures from MR images (for instance, the intervertebral disks (Guerroumi *et al.*, 2019)) and feed the segmented MR data as the input to the CycleGAN model for synthesis. We believe that this approach already provides the model with sufficient surrounding information of the spine so that it can focus only translating the bone structures from MR volumes, thereby easing its task considerably.

APPENDIX I

CONVOLUTIONS

1. The Convolution Operation

The convolution is a technique which changes the intensities of a pixel of an image to reflect the intensities of the surrounding pixels. Let $I(x, y)$ be an input image and let $f(x, y)$ be a function (or, more commonly, a kernel) that is applied to the image. The resulting output is known as a feature map. The 2D convolution in discrete domain is given by:

$$(I * f)(x, y) = \sum_{u=-n}^{n} \sum_{v=-m}^{m} I(x - u, y - v) f(u, v) \tag{A I-1}$$

where, every element of the kernel lies within the range $-n \leq u \leq n$ and $-m \leq v \leq m$. The above equation can be interpreted as follows - Given a 2D image, a kernel (typically, a 3×3 or 5×5 matrix) is slid over the entire image. The overlapping elements of the image and the kernel are element-wise multiplied and the summation of one such multiplication represents one pixel of the new image. The summation of all such multiplications results in a new image.

BIBLIOGRAPHY

Armanious, K. et al. (2019). Unsupervised Medical Image Translation Using Cycle-MedGAN. *CoRR*, abs/1903.03374. Consulted at http://arxiv.org/abs/1903.03374.

Besl, P. J. & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), 239-256. doi: 10.1109/34.121791.

Carter, R. (2021, February, 7). MRI VS. CT Scan; Diagnosing Spine and Neck Injuries and Degenerative Diseases [Web Article]. Consulted at https://josephspine.com/mri-vs-ct-scan-diagnosing-spine-neck-injuries-degenerative-diseases/.

Chan, A. et al. (2019). Does image guidance decrease pedicle screw-related complications in surgical treatment of adolescent idiopathic scoliosis: a systematic review update and meta-analysis. *European Spine Journal*, 29, 694-716.

Chen, Y. & Medioni, G. (1992). Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3), 145-155. doi: https://doi.org/10.1016/0262-8856(92)90066-C.

Chevrefils, C. et al. (2009). Texture Analysis for Automatic Segmentation of Intervertebral Disks of Scoliotic Spines From MR Images. *IEEE Transactions on Information Technology in Biomedicine*, 13, 608-620. doi: 10.1109/TITB.2009.2018286.

Chu, C. et al. (2015). Fully Automatic Localization and Segmentation of 3D Vertebral Bodies from CT/MR Images via a Learning-Based Method. *PLOS ONE*, 10(11), e0143327.

D'Andrea, K. et al. (2015). Utility of Preoperative MRI Co-Registered with Intraoperative CT Scan for the Resection of Complex Tumors of the Spine. *World neurosurgery*, 84.

Daubs, M. D. & Watkins-Castillo, S. I. (2021, February, 5). Scoliosis in Children [Article]. Consulted at https://www.boneandjointburden.org/fourth-edition/iib11/scoliosis-children.

De Silva, T. et al. (2017). Registration of MRI to Intraoperative Radiographs for Target Localization in Spinal Interventions. *Physics in medicine and biology*, 62(2), 684–701. doi: 10.1088/1361-6560/62/2/684.

Delorme, S. et al. (2003). Assessment of the 3-D reconstruction and high-resolution geometrical modeling of the human skeletal trunk from 2-D radiographic images. *IEEE Transactions on Biomedical Engineering*, 50(8), 989-998.

Deng, J. et al. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255. doi: 10.1109/CVPR.2009.5206848.

Egger, J. et al. (2012). Square-Cut: A Segmentation Algorithm on the Basis of a Rectangle Shape. *PLoS ONE*, 7(2). doi: 10.1371/journal.pone.0031064.

Gal, Y. & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of The 33rd International Conference on Machine Learning*, 48(Proceedings of Machine Learning Research), 1050–1059.

Goodfellow, I. et al. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27* (pp. 2672–2680).

Guerroumi, N. et al. (2019). Automatic Segmentation of the Scoliotic Spine from Mr Images. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 480-484.

Hall, E. J. & Brenner, D. J. (2008). Cancer risks from diagnostic radiology. *The British Journal of Radiology*, 81(965), 362-378. doi: 10.1259/bjr/01948454.

Harmouche, R. et al. (2012). 3D registration of MR and X-ray spine images using an articulated model. *Computerized Medical Imaging and Graphics : the Official Journal of the Computerized Medical Imaging Society*, 36, 410-8. doi: 10.1016/j.compmedimag.2012.03.003.

Hemsley, M. et al. (2020). Deep Generative Model for Synthetic-CT Generation with Uncertainty Predictions. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 834–844.

Hiasa, Y. et al. (2018). Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size. *CoRR*, abs/1803.06629. Consulted at http://arxiv.org/abs/1803.06629.

Huang, S.-H. et al. (2009). Learning-Based Vertebra Detection and Iterative Normalized-Cut Segmentation for Spinal MRI. *IEEE Transactions on Medical Imaging*, 28(10), 1595–1605. doi: 10.1109/TMI.2009.2023362.

Isola, P. et al. (2017). Image-to-Image Translation with Conditional Adversarial Networks. 5967-5976. doi: 10.1109/CVPR.2017.632.

Johnson, J. et al. (2016). Perceptual losses for real-time style transfer and super-resolution. *European Conference on Computer Vision*.

Kazeminia, S. et al. (2019). GANs for Medical Image Analysis. *arXiv:1809.06222 [cs, stat]*.

Kendall, A. & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, 30.

Kingma, D. & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Krizhevsky, A. (2012). Learning Multiple Layers of Features from Tiny Images. *University of Toronto*.

Le Cun, Y. et al. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, (NIPS'89), 396–404.

Mahadevan, V. (2018). Anatomy of the vertebral column. *Surgery (Oxford)*, 36(7), 327-332. doi: https://doi.org/10.1016/j.mpsur.2018.05.006.

Mathworks. (2021). Convolutional Neural Networks. Consulted on 2021-02-10 at https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html.

Mitulescu, A. et al. (2001). Validation of the non-stereo corresponding points stereoradiographic 3D reconstruction technique. *Medical & biological engineering & computing*, 39, 152-8. doi: 10.1007/BF02344797.

Miyato, T. et al. (2018). Spectral Normalization for Generative Adversarial Networks. *ArXiv*, abs/1802.05957.

Naga Karthik, E. M. V., Laporte, C. & Cheriet, F. (2021). Three-dimensional segmentation of the scoliotic spine from MRI using unsupervised volume-based MR-CT synthesis. *Medical Imaging 2021: Image Processing*, 11596, 402 – 409. doi: 10.1117/12.2580677.

Neubert, A. et al. (2012). Automated detection, 3D segmentation and analysis of high resolution spine MR images using statistical shape models. *Physics in medicine and biology*, 57, 8357-8376. doi: 10.1088/0031-9155/57/24/8357.

Odena, A., Dumoulin, V. & Olah, C. (2016). Deconvolution and Checkerboard Artifacts. *Distill*. doi: 10.23915/distill.00003.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62-66. doi: 10.1109/TSMC.1979.4310076.

Penney, G. P. et al. (1998). A comparison of similarity measures for use in 2-D-3-D medical image registration. *IEEE Transactions on Medical Imaging*, 17(4), 586-595. doi: 10.1109/42.730403.

Radiopaedia. (2021). Scoliosis. Consulted on 2021-02-05 at https://radiopaedia.org/cases/scoliosis-7.

Rasmussen, C. E. & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rasoulian, A., Rohling, R. N. & Abolmaesumi, P. (2013). A statistical multi-vertebrae shape+pose model for segmentation of CT images. *Medical Imaging 2013: Image-Guided Procedures, Robotic Interventions, and Modeling*, 8671, 184 – 189. doi: 10.1117/12.2007448.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241.

Schwarzenberg, R. et al. (2014). Cube-Cut: Vertebral Body Segmentation in MRI-Data through Cubic-Shaped Divergences. *PLoS ONE*, 9(4), e93389.

Silva, T. D. et al. (2016). 3D–2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch. *Physics in Medicine and Biology*, 61(8), 3009–3025. doi: 10.1088/0031-9155/61/8/3009.

Srivastava, N. et al. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56), 1929-1958. Consulted at http://jmlr.org/papers/v15/srivastava14a.html.

Suzani, A. et al. (2014). Semi-automatic segmentation of vertebral bodies in volumetric MR images using a statistical shape+pose model. In *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling* (vol. 9036, pp. 179 – 184). SPIE.

Ulyanov, D. et al. (2016). Instance Normalization: The Missing Ingredient for Fast Stylization. *ArXiv*, abs/1607.08022.

Wolterink, J. M. et al. (2017). Deep MR to CT Synthesis Using Unpaired Data. *ArXiv*, abs/1708.01155.

Wong, Y.-s. et al. (2019). Is Radiation-Free Ultrasound Accurate for Quantitative Assessment of Spinal Deformity in Idiopathic Scoliosis (IS): A Detailed Analysis With EOS Radiography on 952 Patients. *Ultrasound in Medicine & Biology*, 45(11), 2866–2877.

Yi, X. et al. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58, 101552.

Zhang, Z. et al. (2018, June). Translating and Segmenting Multimodal Medical Volumes With Cycle- and Shape-Consistency Generative Adversarial Network. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, J.-Y. et al. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593 [cs]*, 2242-2251. Consulted at http://arxiv.org/abs/1703.10593. arXiv: 1703.10593.

Zukic, D. et al. (2012). Segmentation of vertebral bodies in MR images. *VMV 2012 - Vision, Modeling and Visualization*, 135-142.