Automatic Assessment of Spinal Deformities in 2D and 3D

Emmanuelle Bourigault

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2025

Abstract

Scoliosis assessment still relies on 2D radiographs and manual measurements by expert clinicians. This thesis proposes deep learning pipelines that automate 2D measurements on dual energy X-ray absorptiometry (DXA) and 3D measurements on Magnetic Resonance Imaging (MRI), unifying 2D and 3D information.

We start by the development of an end-to-end pipeline for 2D scoliosis assessment using DXA scans from the UK Biobank. This foundational work introduces a U-Net style network that segments six body parts including the spine, coupled with an iterative label refinement loop that effectively suppresses segmentation failures. Through cubic spline fitting, we derive precise spine curvature metrics, establishing the spinal geometric framework that underpins the following work in this thesis.

We then shift to a 3D perspective by leveraging volumetric spine MRIs from the UK Biobank. Here, we develop a custom 3D U-Net-like model capable of accurately extracting spine meshes given limited annotation of the spine. This advancement enables comprehensive 3D shape analysis of the spine, capturing critical elements invisible in two dimensions, such as axial rotation and lordosis. The result is a richer set of biomarkers that enhance classification of spine deformity and scoliosis severity.

Building on these advancements, we bridge the gap between imaging modalities by pioneering a technique to predict 3D spine shape directly from a single DXA scan. Our approach employs a vision transformer with a regression head that outputs spine curves from a cropped DXA image, allowing us to reconstruct the full 3D vertebral stack with sub-millimeter accuracy. This image-to-shape model creates opportunities for 3D insights in clinical settings using only low-dose DXA systems.

We also develop an Automated DXA Scoliosis Method (DSM) that directly outputs the maximum angle of the spine, enabling fair comparison with human annotation. This automated approach is validated on manually annotated DXA scans by expert clinicians with strong agreement against expert annotations.

The final stage of our research focuses on enhancing 3D segmentation of the spine to achieve vertebrae-level prediction. We address the challenge of generalizing segmentation to out-of-domain datasets by implementing test-time-adaptation techniques, further improving the robustness and clinical applicability of our methods.

This thesis is submitted to the Department of Engineering Science, The University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is entirely my own work, and except where otherwise stated, describes my own research.

Emmanuelle Bourigault, Apr 2025.

Acknowledgement

First and foremost I am very grateful to my supervisors, Prof. Andrew Zisserman, Dr. Amir Jamaludin and Dr. Timor Kadir for their invaluable advice, and continuous support during my PhD studies. Thank you for the opportunity to work and learn under your guidance. I have learnt how a researcher should think and how research is performed. I would like to thank our collaborators, particularly Prof. Emma Clark for annotating the data to validate my models and for her valuable advice and insights. I would also like to thank Prof. Jeremy Fairbank for introducing me to some key actors in the field of scoliosis research, for helping with data seeking and for insightful discussions. I thank all members of the VGG group for the great company, fruitful discussions and for instauring a positive atmosphere in the lab where it is a pleasure for me to work. I would like to thank especially Ashish Thandavan for promptly answering my questions and for the technical support. I also thank Sarah Clayton and Cassandra Warren for their help. Lastly, I thank my family for their continuous support over the years.

Contents

Chapter 1

Introduction

1.1 Background and Clinical Context

Scoliosis is a three-dimensional deformity characterized by abnormal lateral curvature and axial rotation of the spine. This condition affects patients well beyond adolescence, with epidemiological studies demonstrating that 8 to 32% of adults over forty have spinal curves measuring at least 10° [Schwab et al. 2005; Kilshaw et al. 2019]. The standard quantification method for scoliosis severity is the Cobb angle measurement [Cobb 1948], performed on two-dimensional radiographic images. This technique involves identifying the most tilted vertebrae at the upper and lower boundaries of the curve, drawing lines parallel to their endplates, and measuring the angle formed by perpendicular lines drawn from these endplates. Despite technological advances in medical imaging, this 2D measurement approach remains the clinical standard worldwide due to its established protocols and widespread accessibility.

The manual measurement of Cobb angles introduces substantial variability into clinical assessment and treatment planning. Studies have consistently demonstrated inter-observer differences ranging from 5° to 10° when multiple clinicians measure the same radiograph [Carman et al. 1990; Pruijs et al. 1994; Gstoettner et al. 2007], and even intra-observer inconsistencies when a single clinician remeasures the same image [Morrissy et al. 1990; Ylikoski and Tallroth 1990]. This variability is particularly problematic when measurements fall near critical treatment thresholds (e.g. 20° for bracing consideration or 45° for surgical evaluation). Additionally, the traditional X-ray approach exposes patients to radiation, prompting interest in alternative imaging modalities such as Dual-Energy

X-ray Absorptiometry (DXA). While DXA offers reduced radiation exposure and simultaneous bone density assessment, its typically lower spatial resolution presents additional challenges for accurate curvature measurement, potentially exacerbating the variability problem.

1.2 Research Objectives

This thesis aims to transform the quantification of spinal deformities through computer vision techniques applied to both 2D and 3D imaging. Our primary objective is to develop automated systems that standardize Cobb angle measurements across imaging modalities, eliminating inter-observer variability while maintaining or exceeding the accuracy of expert human assessment. By extending our approach to 3D assessment using MRI data, we address the fundamental limitation of 2D imaging and its inability to fully capture the spine's rotational components that often dictate treatment strategies. Furthermore, we develop models that bridge the gap between lower-cost DXA projections and richer 3D spine representations, providing enhanced anatomical information from minimally invasive scans. The entire pipeline is designed for computational efficiency and scalability to population-level cohorts, enabling rapid translation from research insights to clinical practice while overcoming the challenges of limited expert annotations through semi-supervised learning approaches.

1.3 Clinical Motivation

1.3.1 Scoliosis measurement in clinical practice

Scoliosis is traditionally diagnosed using antero-posterior (AP) X-rays, where an expert clinician measures the lateral deviation of the spine. [Cobb 1948] introduced a manual method for diagnosing scoliosis by measuring the deviation of the spine and vertebral rotation, known as the Cobb angle (see Figure 1.1). This remains the gold standard for diagnosing and quantifying scoliosis severity on AP X-rays, with a Cobb angle greater than 10° considered indicative of scoliosis. However, this method has significant limitations, including variability in measurements, with reported differences ranging from 3° to 10° [Langensiepen et al. 2013]. Additionally, the manual process is time-consuming and restricted to 2D imaging, failing to provide a comprehensive view of

Figure 1.1: **Cobb Angle Measurement.** This manual procedure involves finding the most tilted vertebrae and drawing the lines from the endplates of these vertebrae. The angle at the intersection of these lines is called the Cobb angle [Cobb 1948].

vertebral changes in the sagittal and axial planes.

While this measurement is typically performed on standing AP X-rays, scoliosis can also be assessed using Dual-energy X-ray absorptiometry (DXA) in a lying-down position. The main benefit of DXA is its low radiation meaning it can be safely performed on population at risk such as elderly people. The drawback is the reduced resolution impeding good visualisation of the endplates of the most tilted vertebrae necessary to be able to measure the Cobb angle. [Taylor et al. 2013] adapted the Cobb angle to DXA scans by finding the maximum angle of the spine using body geometry. This consists of drawing the lines through the top of the shoulders and top of the pelvis to identify landmarks to draw the normal spine line (NSL). Then, the apex of the curve is found and the two remaining edges of the triangles are marked by sliding over the NSL at the points of inflection (See Figure 1.2).

DXA is gaining popularity as it not only measures scoliosis but also provides parameters related to body composition enabling further analysis on the relation between curve size and body composition [Clark et al. 2014; Wang et al. 2016; Jamaludin et al. 2020].

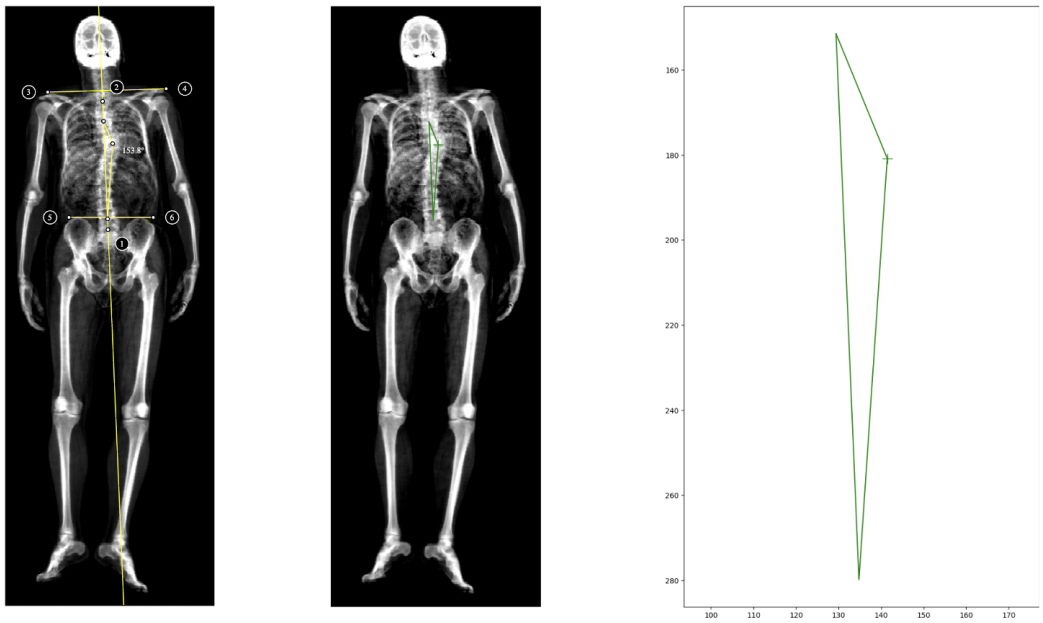


Figure 1.2: **Automated DXA Scoliosis Method (DSM).** Process to find the maximum angle of the spine following the DSM method [Taylor et al. 2013]. The output is a triangle (green) giving the maximum angle of the spine.

1.3.2 Relationship between different planes

Research on the relationship between spinal deformations across the sagittal, axial, and coronal planes remains is in its early stages [Ma et al. 2020; Karam et al. 2022]. The UK Biobank dataset used in this study is notable for its large scale and its focus on an adult cohort. However, most existing research on scoliosis has concentrated on Adolescent Idiopathic Scoliosis (AIS), with comparatively little attention given to scoliosis in adults [Larios et al. 2024; Tang, Walter, et al. 2024; Wang et al. 2021b; Cristante et al. 2021]. In adults, degenerative scoliosis can develop due to the progressive wear and tear of spinal discs. While it is well-established that right thoracic curves are predominantly observed in AIS [Konieczny et al. 2013], similar analyses of spinal curvature in adult scoliosis are scarce.

1.4 Key Research Ideas

Inspired by clinical motivations and the geometry of the spine, we introduce key research ideas in this section that explain the design of our methods given the constraints in terms of limited manual annotation and hardware choice for reproducibility.

Chapter 2

Literature Review

2.1 Datasets

This thesis utilizes multiple large-scale medical imaging datasets spanning different anatomical regions, imaging modalities, and clinical applications. The datasets encompass both dual-energy X-ray absorptiometry (DXA) scans and magnetic resonance imaging (MRI) acquisitions, providing diverse perspectives on human anatomy and pathology. Table 2.1 provides a comprehensive overview of all datasets used throughout this work, while detailed descriptions of each dataset follow below.

Table 2.1: Summary of datasets used across all chapters. Abbreviations: DXA = Dual-energy X-ray Absorptiometry, MRI = Magnetic Resonance Imaging, CT = Computed Tomography

Dataset	Modality	Sample Size	Resolution	Anatomical Region	Annotations	Split (Train:Val:Test)
ALSPAC DXA	DXA	17,389	416×128	Whole body	masks 6-body parts	80:10:10
UK Biobank DXA	DXA	48,384	832×320	Whole body	Scoliosis angles (308)	80:10:10
UK Biobank MRI	MRI	51,761	$501 \times 160 \times 224$	Whole body	Unlabelled	80:10:10
BRATS	MRI	5,880 (1,470 patients)	$240 \times 240 \times 155$	Brain	Tumour regions	80:10:10
BTCV	CT	50	$224 \times 224 \times 85$	Abdomen	13 organs	24:6:20
AMOS	MRI	100	$256 \times 256 \times 125$	Abdomen	11 organs	40:10:50

2.1.1 Whole Body Imaging Datasets

ALSPAC DXA

The Avon Longitudinal Study of Parents and Children (ALSPAC) DXA dataset provides whole-body dual-energy X-ray absorptiometry scans from a longitudinal birth cohort study [Boyd et al. 2013; Fraser et al. 2013]. ALSPAC is a geographically-based UK cohort that recruited pregnant women residing in Avon (South-west England). A total of

14541 pregnancies were enrolled, with 14062 children born. See www.alspac.bris.ac.uk for more information. This dataset includes 7298 children who had DXA scans at the aged 9 research clinic, 5122 who had DXA scans at the aged 15 research clinic, and 4969 who had DXA scans at the aged 17 research clinic. This dataset offers unique insights into bone density development and skeletal health across different age groups, enabling the study of growth patterns and bone health trajectories over time.

UK Biobank DXA

The UK Biobank DXA dataset comprises 48,384 whole-body DXA scans from the UK Biobank cohort, representing one of the largest collections of population-based imaging data available for research [Sudlow et al. 2015]. The dataset is split into training (80%), validation (10%), and testing (10%) sets to ensure robust model development and evaluation. Notably, among the test set, 1,929 have been annotated and among them 308 UK Biobank DXA scans have been expertly annotated with the maximum modified Ferguson angle using the DXA scoliosis method (DSM) as outlined in [Taylor et al. 2013]. This annotated subset provides ground-truth measurements for evaluating the association between predicted spinal curvature metrics and manually annotated angles, enabling validation of our automated scoliosis assessment methods.

UK Biobank MRI

The UK Biobank MRI dataset contains 51,761 full-body MRI scans from more than 50,000 volunteers [Sudlow et al. 2015]. This dataset captures diverse physiological attributes across a broad demographic spectrum, providing unprecedented insights into population health. To ensure consistency and computational feasibility, all UK Biobank MRI scans undergo standardized preprocessing: they are resampled to isotropic voxel spacing and cropped to a consistent resolution. This standardization facilitates large-scale analysis while maintaining anatomical detail necessary for clinical research applications.

2.1.2 Specialized Clinical Datasets

BRATS (Brain Tumor Segmentation)

The Brain Tumor Segmentation (BRATS) dataset represents the largest publicly available collection of brain tumor imaging data, consisting of 5,880 MRI scans from 1,470 patients with brain diffuse glioma [Baid et al. 2021; Menze et al. 2015]. All scans

undergo standardized preprocessing including skull-stripping and resampling to 1 mm isotropic resolution, resulting in uniform images with dimensions of $240 \times 240 \times 155$ voxels. The dataset includes expert clinical annotations delineating three distinct tumor regions: Whole Tumor (WT), Tumor Core (TC), and Enhanced Tumor Core (ET), enabling comprehensive analysis of tumor morphology.

BTCV (Beyond the Cranial Vault)

The Beyond the Cranial Vault (BTCV) dataset focuses on abdominal organ segmentation [Fang and Yan 2020], containing 30 training and 20 testing subjects with annotations for 13 abdominal organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and left adrenal gland. For analysis purposes, the left and right adrenal glands are combined into a single anatomical class. Scans are preprocessed to a consistent resolution of $224 \times 224 \times 85$ voxels, with intensity values scaled to the range [-175, 250] Hounsfield Units (HU), preserving clinically relevant tissue contrast while enabling standardized processing across subjects.

AMOS (Abdominal Multi-Organ Segmentation)

The AMOS dataset, derived from the MICCAI AMOS Challenge, provides 100 abdominal MRI scans split equally between training and testing sets [Ji et al. 2022]. The dataset includes comprehensive segmentations of 11 abdominal organs: liver, spleen, pancreas, kidneys, stomach, gallbladder, esophagus, aorta, inferior vena cava, adrenal glands, and duodenum. All scans are resampled to a consistent resolution of $256 \times 256 \times 125$ voxels, with intensity normalization performed channel-wise to the range [0, 1], ensuring consistent input characteristics for deep learning models while preserving relative tissue contrasts essential for organ delineation.

2.1.3 Dataset Characteristics and Considerations

The datasets employed in this thesis exhibit significant heterogeneity in terms of imaging modalities, population characteristics, and spatial resolutions. This diversity presents both challenges and opportunities for developing robust medical image analysis methods. The UK Biobank datasets provide population-scale data suitable for epidemiological studies, while specialized datasets like BRATS, BTCV, and AMOS offer focused annotations for

specific clinical applications. The variation in spatial resolutions from the high-resolution isotropic MRI data to variable-resolution DXA scans necessitates careful preprocessing and normalization strategies to ensure consistent model performance across different imaging protocols. Furthermore, the demographic diversity represented across these datasets, from the UK-centric ALSPAC and UK Biobank cohorts to the internationally sourced BRATS dataset, enhances the generalizability of developed methods later in this thesis to diverse populations.

2.2 Learning Efficiently with Limited Annotation

Efficient learning with limited annotations refers to the process of training machine learning models, particularly deep learning models, when there is a scarcity of labeled data [Zhu 2005; Goodfellow et al. 2016]. In the domain of medical imaging, such as spine research, annotated data are often scarce due to the expertise required to label images, the time-consuming nature of annotation, and the cost involved in acquiring these labels [Ronneberger et al. 2015; Baur et al. 2017].

In spine research, AI models are developed to assist in tasks such as segmenting vertebrae, predicting spinal conditions, detecting abnormalities, and aiding in surgical planning [Roth et al. 2015; Zhou et al. 2019]. Efficient learning strategies are necessary to leverage the limited labeled data while still achieving high model performance [Cheplygina et al. 2019; Zhu 2005]. AI models in spinal imaging have evolved dramatically from early convolutional neural network (CNN) approaches to sophisticated hybrid architectures incorporating attention mechanisms and transformer-based models. While initial works [Roth et al. 2015; Zhou et al. 2019] demonstrated the feasibility of automated vertebrae segmentation, recent advances have significantly improved both accuracy and clinical applicability.

[Chen et al. 2024] introduced VertXNet, an ensemble method that strategically combines semantic segmentation using U-Net with instance segmentation using Mask R-CNN for vertebral body segmentation and identification from cervical and lumbar X-rays. This dual-pathway approach addresses a fundamental challenge in spinal imaging: differentiating morphologically similar adjacent vertebrae while maintaining anatomical context. The ensemble rule strategy for combining outputs improved overall performance to overcome the limitations of single-model approaches.

The shift toward transformer-based architectures marks a paradigm change in spinal image analysis. VerFormer [Li et al. 2024c] introduces a Vertebrae-aware Vision Transformer that addresses the inherent locality limitations of CNNs through global contextual information processing. The Vertebrae-aware Global Query module specifically targets the challenge of capturing long-range dependencies crucial for accurate spine segmentation. Similarly, SpineHRformer [Zhao et al. 2023] combines HRNet's multi-scale feature extraction with transformer encoders' self-attention mechanisms, demonstrating superior performance in Cobb angle measurement with reduced inter-observer variability.

[Saeed et al. 2023] proposed CHASPPRAU-Net, incorporating cascaded hierarchical atrous spatial pyramid pooling with residual attention mechanisms, achieving notable improvements in handling osteoporotic fractures and spinal anomalies. The model's ability to process varying fracture presentations addresses a critical clinical need, as traditional model-dependent segmentation often fails with vertebral deformities.

The convergence of ensemble methods, transformer architectures, and domain-specific adaptations suggests a trend toward hybrid models that leverage complementary strengths. We introduce below key strategies employed in this work to learn efficiently with limited annotation.

2.3 Active Learning

Active learning is a technique where the model selectively queries for annotations of the most informative and uncertain examples. This method minimizes annotation efforts while maximizing model performance by iteratively updating the training set with high-utility samples. [Settles 2009] first introduced active learning as a means to iteratively refine training datasets by focusing on instances that the model finds most ambiguous. This strategy has been widely adopted in medical imaging tasks to optimize the use of limited annotated data, enabling robust model performance while substantially reducing the manual labeling burden.

[Yang et al. 2017] presented a deep active learning framework that combines fully convolutional network (FCN) and active learning to significantly reduce annotation effort by making judicious suggestions on the most effective annotation areas. Their approach, tested on biomedical image segmentation tasks, demonstrated that state-of-the-art segmentation performance could be achieved using only 50% of training data.

In the context of medical imaging, [Budd et al. 2021] conducted a comprehensive survey on active learning and human-in-the-loop deep learning for medical image analysis, evaluating four key areas including active learning to choose the best data to annotate for optimal model performance. [Smailagic et al. 2018] proposed MedAL, an accurate and robust deep active learning framework for medical image analysis, which achieved 80% accuracy on diabetic retinopathy detection using only 425 labeled images, corresponding to a 32% reduction in the number of required labels.

More recently, [Boehringer et al. 2023] applied active learning to deep learning-assisted segmentation of brain gliomas from MR images, assessing their viability in reducing the required amount of manually annotated ground truth data. Their study demonstrated that active learning approaches could achieve comparable model performance while greatly reducing the time and labor spent on ground-truth training data. [Daniel et al. 2025] proposed RBACA (Replay-Based Architecture for Context Adaptation), a continual active learning framework for medical imaging that employs a rehearsal method to continually learn from diverse contexts while using active learning to select the most informative instances for annotation.

Collectively, these studies underscore the potential of active learning to drive efficient and effective model training across a variety of medical imaging applications, significantly reducing annotation burden while maintaining high performance.

2.4 Semi-Supervised Learning (SSL)

Recent advances in semi-supervised learning have demonstrated significant potential for spine MRI segmentation, achieving performance comparable to fully-supervised methods while reducing annotation requirements [Huang et al. 2023a; Wang et al. 2024a].

[Huang et al. 2023a] proposed SSHSNet, incorporating cross pseudo supervision (CPS) to achieve 96.12% Dice Score for vertebral bodies and 95.07% for intervertebral discs. This two-stage framework effectively utilizes unlabeled data and addresses data imbalance challenges inherent in spine imaging.

Recent work has challenged conventional approaches to uncertainty in SSL. The AC-MT framework [Wang et al. 2022] demonstrated that encouraging consensus on ambiguous regions outperforms traditional uncertainty-based filtering. Similarly, a tripled-uncertainty

guided mean teacher model [Liu et al. 2022] integrated multiple uncertainty measures to improve segmentation reliability.

HD-Teacher [Zhu et al. 2023] introduced a dual 2D/3D mean-teacher framework that captures multi-dimensional information, addressing anisotropic resolution challenges in spine MRI. This approach dynamically combines outputs based on uncertainty scores, achieving superior performance on multi-object segmentation tasks.

[Fiorentino et al. 2024] developed an intensity-based self-supervised domain adaptation method specifically for intervertebral disc segmentation. Their approach learns domain-invariant features across different MRI protocols, demonstrating improved generalization without requiring target domain annotations.

While SSL methods show promising results (>90% Dice scores), several limitations persist: (i) limited multi-center validation [Pang et al. 2022], (ii) computational overhead from hybrid architectures; (iii) lack of standardized benchmarks beyond the SPIDER challenge [Graaf et al. 2024].

2.5 Transfer Learning and Pretrained Models

Transfer learning has emerged as a valuable approach in spinal imaging analysis, significantly reducing dependency on large labeled datasets. [Oquab et al. 2014] pioneered the concept of fine-tuning deep convolutional networks for specific tasks with limited labeled data, establishing a foundation widely adopted in spinal image classification and segmentation. [Shin et al. 2016] demonstrated that adapting pre-trained CNN architectures substantially improves detection and segmentation accuracy across various medical imaging modalities. Recent systematic reviews [Morid et al. 2021; Kim et al. 2022] confirm transfer learning's prevalence, with 92% of deep learning spine imaging studies developing new models while leveraging pre-trained weights [Constant et al. 2023]. The emergence of foundation models such has revolutionized this field. SpineFM [Simons et al. 2025] leverages the Medical-SAM-Adaptor to achieve 97.8% and 99.6% vertebrae identification rates, while MA-SAM [Fan et al. 2024] demonstrates multi-atlas guided segmentation without manual annotation. [Mazurowski et al. 2023] evaluated SAM's performance across 19 medical imaging datasets, revealing highly variable results for spine MRI (IoU=0.1135) compared to other modalities. Advanced architectures now combine transfer learning with specialized spine-specific modifications. [Qadri

et al. 2023] proposed patch-based deep learning with stacked sparse autoencoders, while [Cheng et al. 2021] developed two-stage Dense-U-Net achieving 0.953 Dice coefficient. [Meng et al. 2022] introduced graph optimization with anatomic consistency cycles for handling pathological cases. These advances, documented in comprehensive reviews [Shi et al. 2025; Qu et al. 2022], collectively demonstrate that transfer learning and foundation models offer robust solutions for advancing spine research through efficient utilization of scarce annotated data.

2.6 3D Shape Analysis of Scoliosis

Until now, the vast majority of scoliosis research has focused on 2D shape analysis of the spine, but not in 3D at a large scale. Limitations of 2D spine analysis arise particularly in classifying curve shape [Zhang et al. 2019]. Indeed, deviations are not limited to the coronal plane, they include twisting of the spine in multiple directions [Saito and Tanaka 2020]. Consequently, the axial and sagittal planes have been largely ignored in previous studies [Ma et al. 2020]. Recent work has started to address these limitations by examining 3D curves in scoliosis using EOS images, a low-dose full-body X-ray technology with accurate three-dimensional reconstruction [Hu et al. 2021; Gajny et al. 2019; Ilharreborde et al. 2011; Karam et al. 2022; Smith et al. 2019]. However, these studies are typically limited to small sample sizes (n = 100) and focus predominantly on adolescent idiopathic scoliosis (AIS), as opposed to degenerative scoliosis observed in larger cohorts such as the UKBiobank. Although our work employs MRIs, it is worth noting that most investigations using spinal MRIs have concentrated on non-scoliosis spinal disorders placing more emphasis on the segmentation of vertebral bodies and intervertebral discs rather than on the spine as a whole [D'Andrea et al. 2021; Jones et al. 2018].

2.7 Thesis Outline and Contributions

Chapter 3: Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach. We begin with the development of an automated pipeline for 2D scoliosis analysis using DXA scans [Bourigault et al. 2022]. In this chapter, we focus on segmenting the spine, refining the segmentation labels through iterative training, and computing spine curvature metrics. This foundational work sets

the stage for exploring more complex representations of spinal deformity.

Chapter 4: 3D Shape Analysis of Scoliosis. Building on the 2D analysis, we transition to a 3D perspective by leveraging MRI data from the UK Biobank [Bourigault et al. 2023]. Here, we explore methods for assessing scoliosis in three dimensions, providing richer anatomical insight and enabling more comprehensive characterization of spinal curvature.

Chapter 5: 3D Spine Shape from 2D DXA. Recognizing the value of combining the accessibility of DXA scans with the detail of 3D representations, we introduce an image-based regression model that estimates 3D spine shape from a single 2D DXA scan [Bourigault et al. 2024b]. This chapter bridges the gap between imaging modalities and highlights the potential of predictive modeling to enhance clinical workflows.

Chapter 6: Automated DXA Scoliosis Method (DSM). We design and validate an automated method for computing the spine's maximum angle on a subset of the UK Biobank dataset (n = 1,929) annotated by human experts. This validation relies on the segmentation and spline fitting techniques described in **Chapter 3**. We also extend this work by analyzing curve patterns to define phenotypic characteristics of scoliosis, offering a data-driven approach to subclassifying spinal deformities.

Chapter 7: UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation. We improve the 3D spine segmentation labels from chapter 4 and treat current drawbacks in medical imaging segmentation impeding good generalisation to external datasets such as small scale dataset used for training, difference in imaging protocols, and modality difference [Bourigault et al. 2025]. We leverage one of the largest oublicly available dataset of more than 51K MRI from the UK Biobank [Sudlow et al. 2015] and segmentation labels from TotalVibeSegmentator [Graf et al. 2024] that we filtered with our custom body filtration method. We employ state-of-the-art segmentator model on this filtered dataset and use entropy test-time adaptation which showed good performance on abdomen organs CT and MRI, and brain MRI public datasets.

Chapter 8: Summary and Extensions. Finally, we provide a summary of the key findings across all chapters. We reflect on the contributions made throughout this work, and we discuss potential future directions for extending and applying these methods. In particular, we discuss the feasibility and robustness of our approach in real-world population settings, enabling population-level scoliosis research.

2.8 Publications

Chapter 3 to **7** each contains a research paper. They all have been peer-reviewed and accepted for publication at a conference, with the exception of **Chapter 6** which is under submission. We make no modifications to the published papers except for formatting changes. For each paper, we also provide a statement of authorship in **Appendix** A. The papers included in the thesis are listed below.

Chapter 3: "Scoliosis measurement on DXA scans using a combined deep learning and spinal geometry approach" **Emmanuelle Bourigault**, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. In Medical Imaging and Deep Learning (MIDL), 2022.

Chapter 4: "3D Shape Analysis of Scoliosis" **Emmanuelle Bourigault**, Amir Jamaludin, Emma M. Clark, Jeremy Fairbank, Timor Kadir, and Andrew Zisserman. Shape in Medical Imaging (ShapeMI), 2023.

Chapter 5: "3D Spine Shape Estimation from Single 2D DXA" **Emmanuelle Bourigault**, Amir Jamaludin, and Andrew Zisserman. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2024.

Chapter 6: "Automated DXA Scoliosis Method" **Emmanuelle Bourigault**, Amir Jamaludin, Emma M. Clark, Jeremy Fairbank, Timor Kadir and Andrew Zisserman. *Under submission.*

Chapter 7: "UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation" **Emmanuelle Bourigault**, Amir Jamaludin, and Abdullah Hamdi. International Conference on Computer Vision (ICCV), 2025.

Chapter 3

Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach

This paper was published in the proceedings of the Medical Imaging and Deep Learning Conference (MIDL), 2022 [Bourigault et al. 2022].

In the first study of this PhD thesis, we extend previous research on two-dimensional curvature analysis of DXA scans from [Jamaludin et al. 2019a]. The primary objective was to develop an automated approach to measure scoliosis in adults using DXA scans from the UK Biobank. Adult scoliosis has received considerably less attention than Adolescent Idiopathic Scoliosis (AIS). Building upon the work of [Jamaludin et al. 2019a], which demonstrated that curvature can serve as a reliable continuous-scale metric for scoliosis measurement, this study revisits and enhances that methodology.

Traditionally, scoliosis has been assessed using the Cobb angle; however, the Cobb method is limited by significant inter-observer variability, typically ranging between 5° and 10°. To address these discrepancies and improve upon previous findings, this study introduces two major enhancements. First, we propose a spline-based geometric representation of the spine for curvature measurement instead of using the midpoint of the spine segmentation. This modification reduces the noise associated with midpoint predictions, thereby increasing agreement with expert human annotations. Second, we implement a pseudo-labelling strategy for segmentation to mitigate the domain gap between the Avon Longitudinal Study of Parents and Children (ALSPAC) and UK

Biobank DXA scans.

These methodological improvements not only refine the measurement of scoliosis in adult populations but also lay a robust foundation for the subsequent studies presented in this thesis.

Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach

Emmanuelle Bourigault Amir Jamaludin Timor Kadir

Andrew Zisserman

VGG, Department of Engineering Science, University of Oxford

emmanuelle,amirj,az@robots.ox.ac.uk

Abstract

We propose improvements to an automated method for scoliosis measurement [Jamaludin et al. 2019a]. Our main novelty is the use of a spline to better model the curve of the spine, and we employ pseudo-labelling approach by iterative training of the segmentation model to mitigate the domain gap when adapting to a new dataset. We obtain promising results with a good fit of our smoothed curve to approximate the spinal midpoints in severe scoliosis cases, and obtain good agreement against human ground-truth. This work is relevant for improving the severity grading of scoliosis and potentially aiding in the treatment management of scoliosis.

Scoliosis measurement, Image Segmentation, Deep Learning, Computational Geometry

3.1 Introduction

Scoliosis is a spinal deformation affecting between 2 to 3% of the population worldwide. Most automated methods for scoliosis measurement focus on Adolescent Idiopathic Scoliosis (AIS) but in this work our aim is to measure scoliosis in adults using DXA scans in the UK Biobank. We use the work of [Jamaludin et al. 2019a], which showed

curvature is a reliable method to measure scoliosis, as a base and propose several adjustments. Note that the model built by [Jamaludin et al. 2019a] was validated on a completely different dataset, ALSPAC (The Avon Longitudinal Study of Parents and Children), from the UK Biobank with 2 main differences: (i) ALSPAC scans are lower in resolution (5$\times$ lower) compared to the UK Biobank, and (ii) ALSPAC participants are all adolescents while the UK Biobank is predominantly made up of adult participants. We make two contributions: (i) spline geometric representation and curvature measurement of the spine, rather than the mid-point of the spine segmentation used in [Jamaludin et al. 2019a]; and (ii) pseudo-labelling the segmentation to address the domain gap between ALSPAC and UK Biobank DXA scans.

3.2 Methods

3.2.1 Dataset

The dataset comprises 48,384 whole body DXA scans from the UK Biobank split into 80:10:10 for training, validation and testing. Among the test set, 308 UK Biobank DXA scans have been annotated with the maximum modified Ferguson angle using the DXA scoliosis method (DSM) as outlined in [Taylor et al. 2013]. We use this test set to evaluate the association between our predicted curvature and manually annotated angle.

3.2.2 Pseudo-labelling

The purpose of our pseudo-labelling is to bootstrap from the ALSPAC model to the higher resolution UK Biobank images. The initial labels are obtained as follows: UK Biobank DXA scans are downsampled to the ALSPAC resolution, the ALSPAC model is applied and segmentation masks are upsampled. We first train our model on 80% of the training set and incrementally add 10% of the training data with each iteration. Further, the generated pseudo-labels are combined with the original labels to re-train the model.

3.2.3 Segmentation Task

We use as a baseline the ALSPAC model with input resolution (416×128). Then, we train our higher resolution segmentation model which is a U-Net like architecture [Ronneberger et al. 2015] depicted in Figure 3.1. We feed into the network the normalised

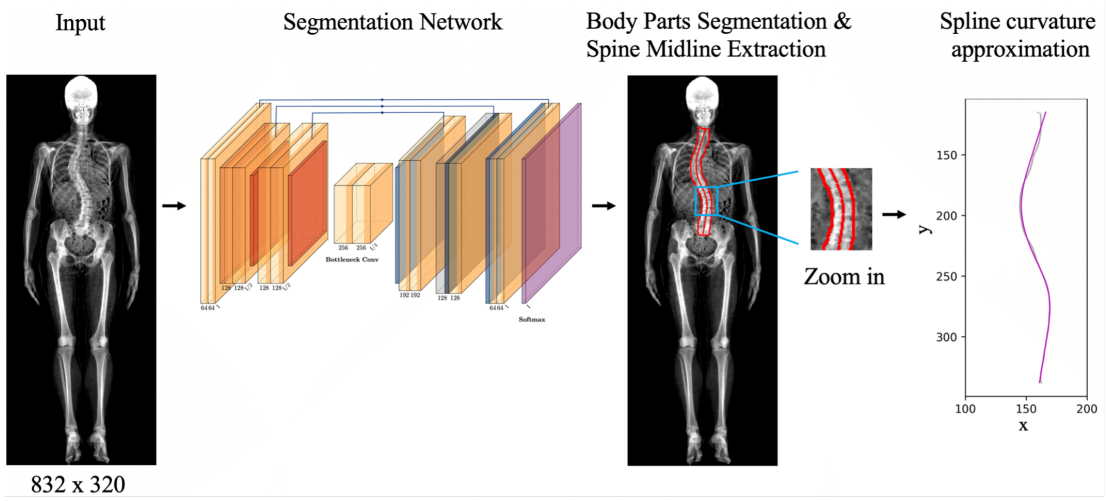


Figure 3.1: **Pipeline of the automated scoliosis measurement method.**

UK Biobank DXA scan as tensor $(1 \times 832 \times 320)$. The network attempts to predict the associated ground-truth masks for 6 body parts: head, spine, pelvis, cavity, left leg and right leg. The decoder outputs probability maps of size $(6 \times 832 \times 320)$. At each row of the spine probability map, the weighted arithmetic mean of the probability and the indices of the scores is calculated to be the predicted midpoint.

3.2.4 Spline Fitting Task and Curvature Measurement

The extracted spinal midpoints from the spine probability maps can be slightly noisy especially near cases with unclear boundaries of the spine (see Figure 3.1). Hence, we propose a smoothed piecewise cubic spline that locally approximates the spinal midpoints and effectively handles the noise in the data. At each segment, a cubic polynomial is used. Constraints on continuity and smoothness at the knots are added. Piecewise cubic splines are determined by minimising the sum of the weighted squared residuals between the fitted spline and data points [Ezhov et al. 2018]. The number of interior knots, k, controls the goodness of fit. If k is too small, the spline will be too smooth, if k is too large, the spline will capture the noise.

We compute curvature at each point along the curve with the traditional formula for plane curves. Then, we compare the maximum curvature with human annotated angles. We assume that the maximum curvature is proportional to the maximum angle.

3.3 Results

3.3.1 Segmentation Task

The Root Mean Squared Error (RMSE) is computed as the root mean squared nearest Euclidian distance between predicted to manual contour points. The lower the RMSE, the better is model performance. Table 3.1 shows the improvement in segmentation accuracy (Intersection over Union (IoU) = 0.94, RMSE = 0.83) with pseudo-labelling compared to the baseline model (IoU = 0.89, RMSE = 1.18). The IoU increases by 2% while the RMSE decreases by 0.12 at sub-pixel level precision for the iteration of the pseudo-labelling.

Models	**IoU**	**RMSE**
ALSPAC model on the UK Biobank	0.86	1.80
Model trained on UK Biobank with pseudo-labels no iteration	0.89	1.18
Model trained on UK Biobank with pseudo-labels with iterations	0.94	0.83

Table 3.1: Table of model segmentation performance with pseudo labelling

3.3.2 Curvature to Angle Correspondence

Comparison of maximum curvature to maximum human annotated angle suggests a good agreement (Pearson's r = 0.92) with a tendency for our method to underestimate the angle for angles below 20° (see Figure 3.2).

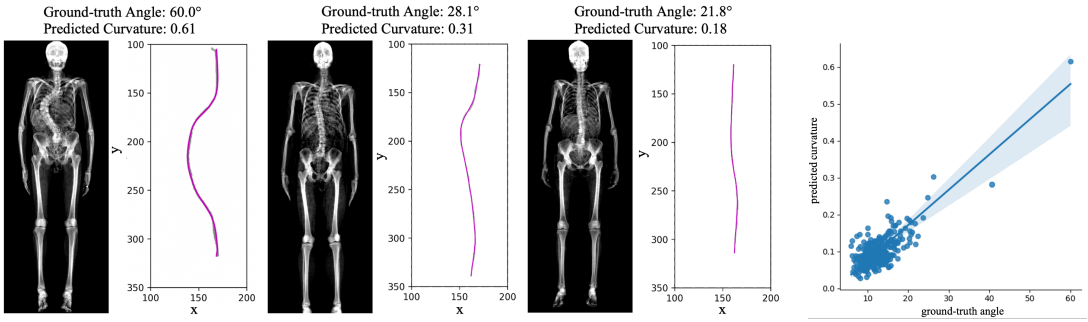


Figure 3.2: **Comparison between predicted curvature and human annotated angles.**

3.4 Conclusion

We have implemented a model for the measurement of scoliosis by leveraging one of the largest DXA datasets available. The benefit of our pseudo-labelling in training is 2-fold, it improves the accuracy of the segmented spine compared to our baseline model, and it

overcomes the low availability of human annotated data sets. This work has implications for improving scoliosis detection and severity measurements in adolescents and adults.

Chapter 4

3D Shape Analysis of Scoliosis

This paper was published in the proceedings of the Shape MI Workshop of the Medical Image Computing and Computer Assisted Intervention Conference (MICCAI), 2023 [Bourigault et al. 2023].

This study builds upon previous Chapter 3 on 2D scoliosis analysis by incorporating 3D MRI data to provide a more comprehensive understanding of spinal deformities. We introduced the concept of axial rotation, defined as the angle between the centroids of the spine and vertebral canal. We established that curvature metrics correlates well with 3D measurements, thereby validating the 2D observations while revealing additional nuances such as the generally lower curvature of the vertebral canal compared to the spine in normal cases, and demonstrating moderate to strong correlations between axial rotation and maximum coronal curvature. This integrated approach not only reinforces the reliability of traditional 2D assessments but also lays the groundwork for advanced unified 3D classification methods that could enhance diagnostic precision by bridging the gap between 2D and 3D spine characterisations.

3D Shape Analysis of Scoliosis

Emmanuelle Bourigault
VGG, University of Oxford

Amir Jamaludin
VGG, University of Oxford

Emma M. Clark
University of Bristol

Jeremy Fairbank
Nuffield Department of Orthopaedics, University of Oxford

Timor Kadir
Plexalis Ltd

Andrew Zisserman
VGG, University of Oxford

emmanuelle,amirj,az@robots.ox.ac.uk

Abstract

Scoliosis is typically measured in 2D in the coronal plane, although it is a three-dimensional (3D) condition. Our objective in this work is to analyse the 3D geometry of the spine and its relationship to the vertebral canal. To this end, we make three contributions: first, we extract the 3D space curve of the spine automatically from low-resolution whole-body Dixon MRIs and obtain coronal, sagittal and axial projections for various degrees of scoliosis; second, we also extract the vertebral canal as a 3D curve from the MRIs, and examine the relationship between the two 3D curves; and third, we measure the angle of rotation of the spine and examine the correlation between this 3D measurement and the 2D curvature of

the coronal projection. For this study, we use 48,384 MRIs from the UK Biobank.

MRI, Spine Geometry, 3D/2D Correspondences

4.1 Introduction

Scoliosis is defined as a lateral deformation of the spine in the coronal plane, usually manually diagnosed on anteroposterior (AP) X-rays, by measuring the Cobb angle, where an angle over 10 degrees is considered scoliotic [Cobb 1948]. More recently, it has been shown that scoliosis can also be diagnosed from DXA (Dual-energy X-ray Absorptiometry) scans, which are less costly and involve a 10 times lower radiation dose than conventional X-rays [Taylor et al. 2013]. However, both X-rays and DXAs do not capture the complex 3D deformation of the spine [Illés et al. 2010]. The convenience of using coronal radiographs to measure scoliosis has meant that the axial and sagittal planes have been widely disregarded.

In this work, we explore scoliosis in 3D by analysing the 3D shape of the space curve of the spine, and its relationship to the 3D space curve of the vertebral canal. For this study, we use the Dixon MRIs available in the UK Biobank. We segment both the spine and the vertebral canal in axial slices. These segmentations allow us to extract 3D curves for the spine and canal, as illustrated in Figure 4.1.

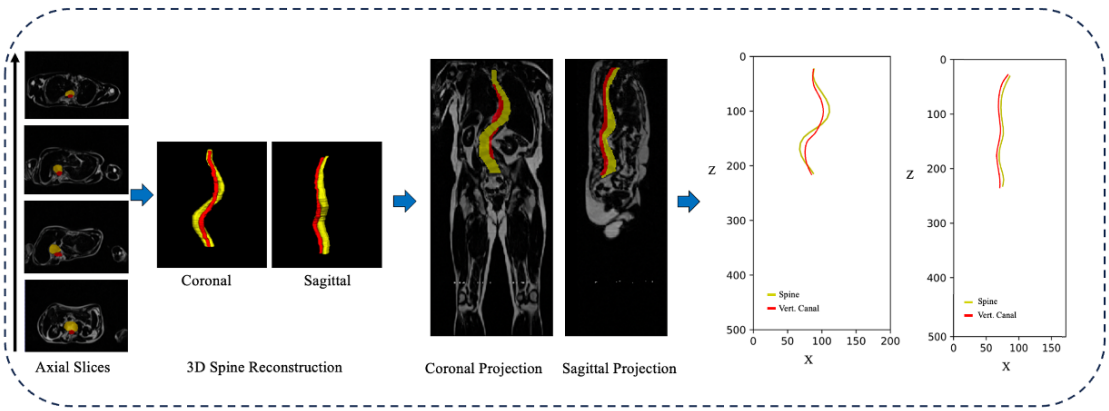


Figure 4.1: **Overview of the geometry pipeline.** The spine (yellow) and the vertebral canal (red) are segmented in each axial slice. The centroids of the spine segments over all axial slices form a 3D space curve (similarly for the canal). The space curve is projected onto the coronal and sagittal planes, and a 2D spline curve fitted to the projected points. Curvature and angles are computed from the spline curve.

Our objective is to study how the 3D spine curve deforms for a scoliotic spine, and also how the vertebral canal adapts to scoliosis. We analyse the 3D spine curve by projecting

it onto coronal, sagittal and axial planes, and determine the severity of scoliosis on the coronal plane. It is worth noting that the MRIs from the UK Biobank are uniquely suitable for scoliosis measurement in 3D as there exists an established scoliosis measurement on the paired MRI to 2D DXA for the coronal projection which serves as our point of reference [Bourigault et al. 2022; Jamaludin et al. 2019a].

We then investigate the relationship between the spine and vertebral canal curves on the three planes, and also measure the deviation between the two curves. In addition, we measure the curvature of the coronal projection and the angle of axial rotation of the spine; and investigate their relation. Section 4.2 outlines our method for extracting the geometry of the spine and vertebral canal from MRIs, and describes the measures we use for the analysis of the geometry. Then, Section 4.3 describes the dataset, and presents the results of the analysis, together with several visualizations of the geometry. Finally, Section 4.4 summarises the findings and the implications of this research.

4.1.1 Related Work

Research on the relationship between deformations on the sagittal, axial and coronal planes is still in its early phases [Ma et al. 2020; Karam et al. 2022].

The UK Biobank dataset used in this paper is of adults. However, most work on scoliosis focuses on adolescent idiopathic scoliosis (AIS), while scoliosis in adults has been relatively unexplored in past literature. Grown adults can develop degenerative scoliosis as a result of wear and tear on the discs of the spine. It has been shown that the right thoracic curves are predominant in AIS [Konieczny et al. 2013] but this kind of shape analysis of the spine in adult scoliosis is rare.

To date, the vast majority of scoliosis research has focused on 2D shape analysis of the spine, but not in 3D at a large scale. Limitations of 2D spine analysis arise particularly in classifying curve shape. Indeed, deviations are not limited to the coronal plane. They include twisting of the spine in multiple directions [Rockenfeller and Müller 2022]. The closest work to ours is by [Pasha 2018] in which they look at 3D curves in scoliosis. The main differences between their work and ours are: they used EOS which is quite a niche imaging modality compared to MRIs, they focused on AIS as opposed to degenerative scoliosis, and the number of samples is small (n=103).

Though we use MRIs in our work, it is worth noting that most works on spinal MRIs

focus on non-scoliosis spinal disorders and as such put more emphasis on segmenting the vertebral bodies and discs individually rather than the spine as a whole [Khalil et al. 2022; Jamaludin et al. 2016; Windsor et al. 2020; Ma et al. 2020; Karam et al. 2022].

4.2 The 3D Geometry of the Spine and Vertebral Canal

It is essential for the 3D geometrical analysis of scoliosis that we capture the whole shape of the spine. To this end, we segment the two main structures that can be seen in the axial Dixon MRIs: these are (i) the “spine" itself, which is comprised of the vertebral bodies and the intervertebral discs, and (ii) the “vertebral canal", which is the space occupied by the spinal cord and filled with cerebrospinal fluid. We do this segmentation on a per-slice basis for each axial image in a given scan volume. The centroid (a 2D point) of each segmentation can now be extracted from each slice and stacked vertically according to axial slice numbering, and spaced appropriately with the axial slice thickness, for a given volume.

A spline curve can then be fitted in 3D (or to the 2D projections), to smooth out the noise in the measurements. The full implementation details are given in Appendix B, and the process is summarised in Figure 4.1. The spine segmentation gives us the 3D spine curve and the vertebral canal segmentation gives us the 3D vertebral canal curve. These 3D space curves can be projected to the coronal (X,Z), the sagittal (Y,Z), as well as the axial (X,Y) plane. Figure 4.2 shows an example of two curves in our dataset rotating in space.

4.2.1 Measuring the Deviation of the Curves

Now that the 3D curves of the spine and the vertebral canal have been extracted, we can then proceed to the analysis of the curves. For a given normal spine of maximum angle 8° or less based on manually annotated sample by expert clinicians, it can be observed that two curves should overlap when projected onto the coronal plane and should be parallel in the sagittal plane. As such, the simplest measurement that is indicative of how far away from the norm a given pair of curves are is to measure the deviation between these two curves. Simply put, to measure the distance, ‘d’, between the two curves we can simply compute the vector joining each point of the spine (x_{spine}, y_{spine}) to the vertebral canal curve (x_{canal}, y_{canal}) (see Figure 4.3). If we project these vectors to the coronal plane, then the sum of their magnitudes measures the ‘deviation’ between the two curves.

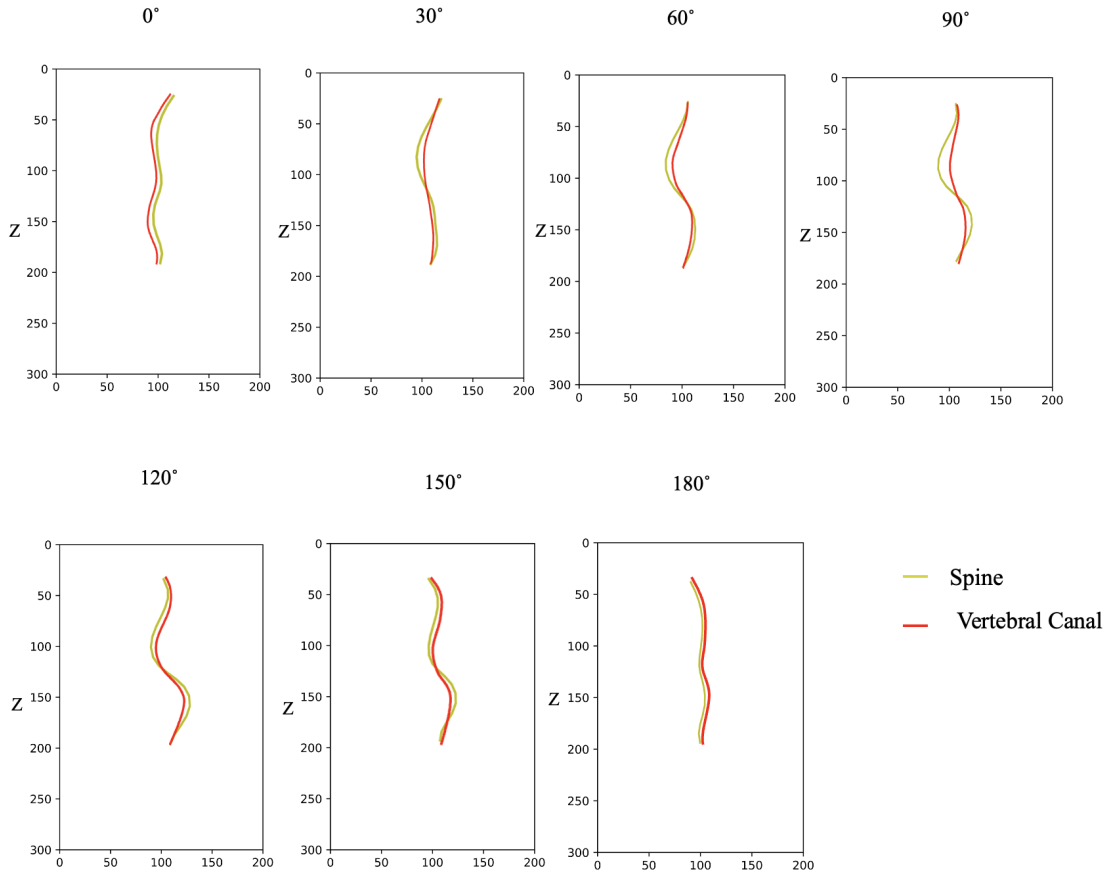


Figure 4.2: **Spine and Canal 2D Projections for every 30 degree of rotation for a severe 'S shape' spine.** This example is in Figure 1. The 0 degree projection corresponds to the sagittal projection, and the 90 projection to the coronal projection.

For a normal ideal spine, the 'deviation' will be zero in the coronal plane (since the spine and vertebral canal will project on top of one another), and in the sagittal plane, the point-wise difference between the two curves will have a set but constant 'deviation'.

For a given pair of spine and vertebral canal curves, we compute the point-to-point distance in the axial plane:

$$\delta_{\text{spine-canal}} = \sum_{i=1}^{N} \sqrt{\left(x_i^{\text{spine}} - x_i^{\text{canal}}\right)^2 + \left(y_i^{\text{spine}} - y_i^{\text{canal}}\right)^2} \tag{4.1}$$

where i is the slice index and N is the total number of axial slices containing the spine and canal for a given scan.

Then, we can obtain the maximum deviation by taking the maximum of the point-wise distances of the spine-canal deviation 4.1. For a normal spine, the maximum deviation will be zero on the coronal plane. For the sagittal plane, a normal spine has inward curvature (lordosis) for the cervical and lumbar sections, and outward spinal curvature

(kyphosis) for the thoracic section. Sagittal malalignment is as an exaggeration or deficiency of the normal lordosis or kyphosis curves. Therefore, we measure how parallel the spine and canal curves are on the sagittal plane by taking the standard deviation of the spine-canal deviations.

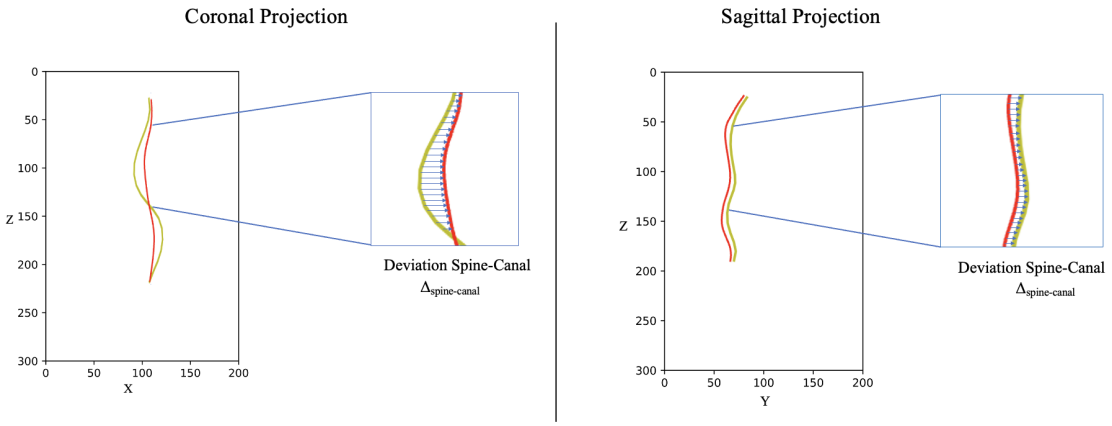


Figure 4.3: **Measurement of deviation between spine (yellow) and canal (red) curves.** Shown is a coronal projection (left) and sagittal projection (right) and a zoom in on the maximum coronal curvature point.

4.2.2 Curvature of the Spine Curve

The cubic spline is continuous everywhere, as are its first and second derivatives. This is sufficient to determine the curvature κ with the standard mathematical formula:

$$\kappa = \frac{(y'' x' - x'' y')^{\frac{3}{2}}}{(x'^2 + y'^2)} \tag{4.2}$$

where:

$$\begin{aligned} x' &= \frac{dx}{dt} &&= \text{First derivative of } x \text{ with respect to parameter } t \\ y' &= \frac{dy}{dt} &&= \text{First derivative of } y \text{ with respect to parameter } t \\ x'' &= \frac{d^2x}{dt^2} &&= \text{Second derivative of } x \text{ with respect to parameter } t \\ y'' &= \frac{d^2y}{dt^2} &&= \text{Second derivative of } y \text{ with respect to parameter } t \\ t &&&= \text{Parameter along the spine curve (normalized position)} \end{aligned}$$

For the results in Section 4.3 the maximum absolute curvature in the coronal plane is

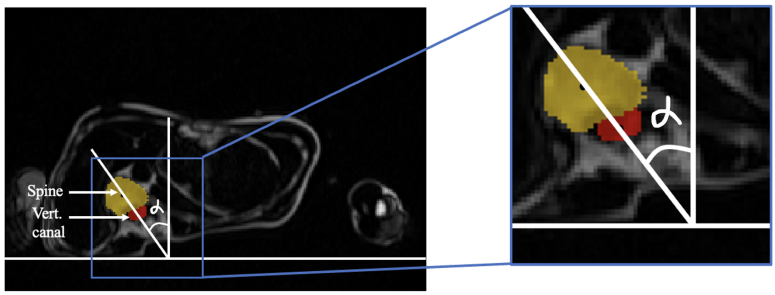


Figure 4.4: **Angle of axial rotation**. The angle of axial rotation, α, is the angle between the line through the centroids of the spine (yellow) and vertebral canal (red), and the vertical direction.

used to define three classes of scoliosis severity (normal, mild, severe) according to thresholds obtained on a set of 2K DXA scans annotated for Cobb angles. The threshold for scoliosis is $|\kappa| = 0.083$, mild scoliosis is: $0.083 < |\kappa| \leq 0.118$; and $|\kappa| > 0.208$ is severe scoliosis.

4.2.3 Angle of Spinal Axial Rotation

Aside from measuring the deviation of the two curves, we can also evaluate the lateral shift of the spine relative to the vertebral canal by measuring the angle of rotation. This is done by using two landmarks: the centroid of the spine and the centroid of the vertebral canal (see Figure 4.4). The angle between the line through these centroids and the vertical is the axial rotation (under the assumption that the patient is lying on their back). Note, there are several definitions of the angle of axial rotation. They all rely on measuring the relative positions of anatomical landmarks such as the pedicles, vertebral body, and spinous processes. We use a similar approach to that of [Aaro et al. 1978] and [Ho et al. 1993], but choose to detect the vertebral canal as a landmark on our axial slices as it is continuous throughout the spine.

4.3 Results and Discussion

In this section, we compare the 2D projected curvature in relation to the 3D spine. We investigate how the canal curve varies with respect to the spine in Subsection 4.3.2. And in Section 4.3.3, we analyse the coronal and sagittal curvatures and their relation to the angle of maximum axial rotation.

4.3.1 Dataset

Our dataset is comprised of 48,384 whole-body MRIs from the UK Biobank, a large open-access medical dataset with scans from more than 500,000 volunteers [Sudlow et al. 2015]. MRIs in the UKBiobank are of much lower resolution than standard clinical scans. Scans are resampled to be isotropic and cropped to a consistent resolution ($501 \times 156 \times 224$). The dataset is split into 80:10:10 for training (38,707), validation (4,838), and testing (4,839) for the segmentation task. 250-200-200 MRI scans are annotated for train-validation-testing for spine and vertebral canal for the baseline segmentation model and checked by an expert clinician. A part of the testing set (1,929) has been annotated by experts for Cobb angles using a modified Ferguson method in whole-body DXA scans as described in [Taylor et al. 2013]. We use this annotated set to define the threshold for scoliosis in our experiment; otherwise this test set is unused in the training of our pipeline. Appendix B gives details of the segmentation.

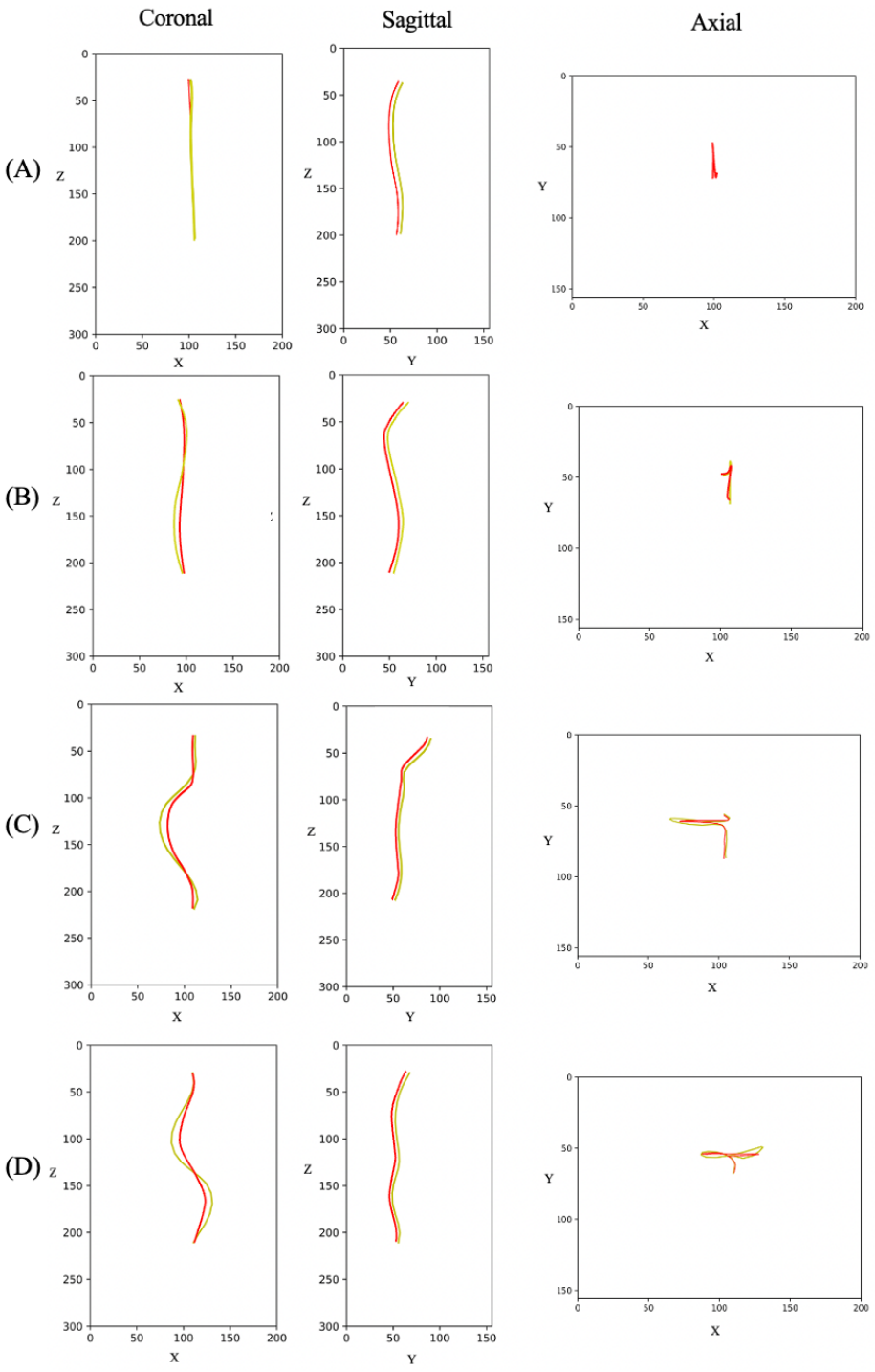


Figure 4.5: **Comparison of coronal, sagittal and axial 2D projections from 3D curve for normal (A), mild (B) and severe C shape (C) and severe S shape (D) scoliosis cases.** Spine curve is in yellow and vertebral canal curve in red.

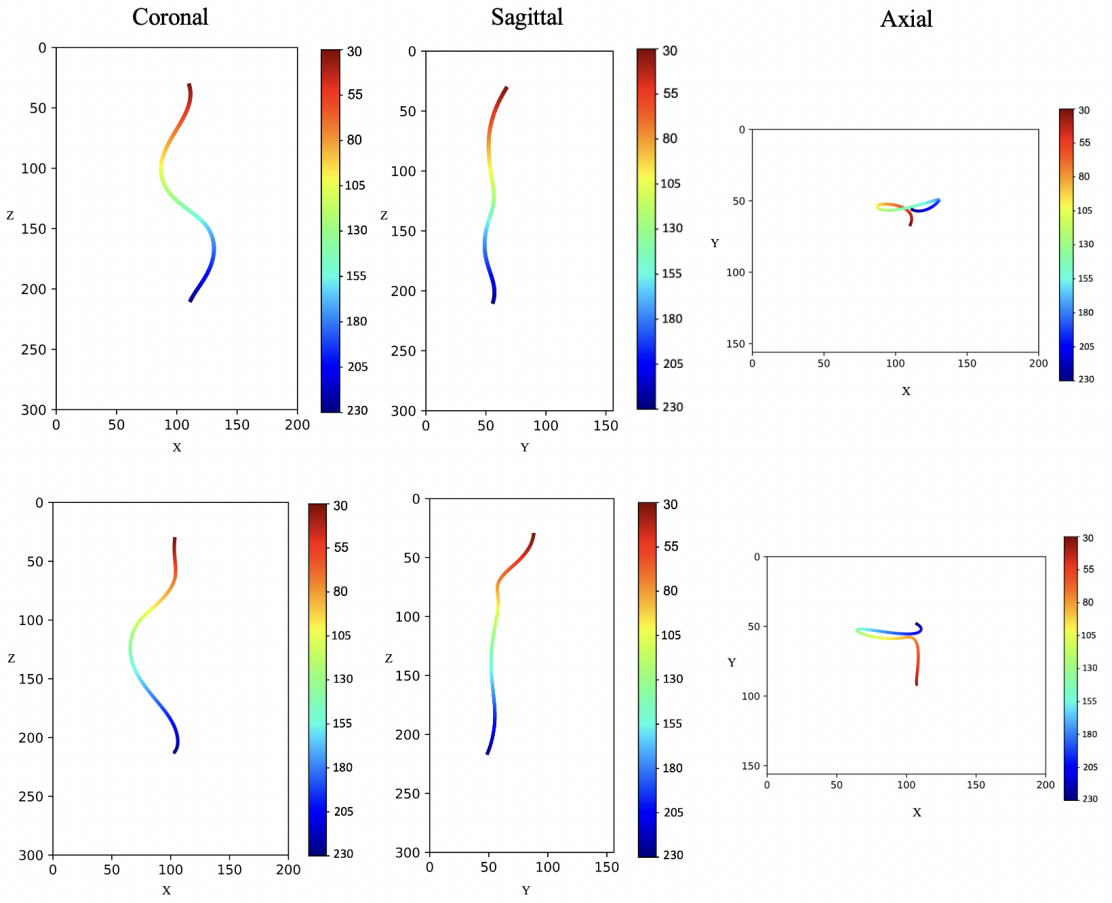


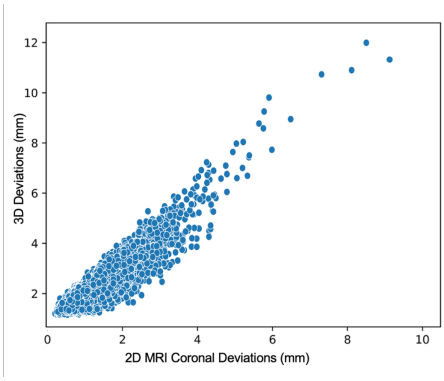
Figure 4.6: **Comparison of coronal, sagittal and axial 2D projections from 3D curve for a severe S shape (top), and a severe C shape (bottom) scoliosis case.** The axial curves (3rd column) are more challenging to interpret. Spines are colour-coded on the z axis to visually indicate the order of the curve in the other projections.

4.3.2 Geometry of the Spine: Deviation of the Spine and Vertebral Canal

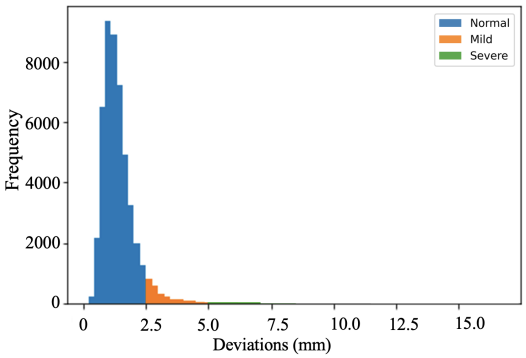
For a normal case, the spine and canal overlap in coronal, and are at a constant separation in sagittal (see Figure 4.5). By comparing the curves of the spine and canal for normal versus scoliosis cases, we observe that the curves on coronal for scoliosis cases no longer overlap. We also observe that the vertebral canal is less curved than the spine suggesting that it deforms less than the spine. On the sagittal plane, the curves straighten from normal to scoliosis cases (see Figures 4.5 and 4.6).

We study how the 3D deviation measurements relate to 2D. The results confirm a strong correlation in deviations between the spine and vertebral canal in 2D coronal and 3D curves (see Figure 4.7(a)). The threshold for scoliosis is $|\kappa| = 0.083$. We define mild scoliosis as: $0.083 < |\kappa| \leq 0.118$; and $|\kappa| > 0.208$ for severe scoliosis.

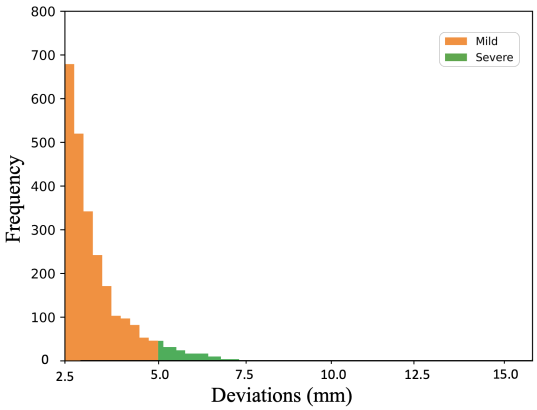
Distribution of spine-canal deviations can be discretised according to scoliosis severity (see Figure 4.7(*b*)). This suggests that spine-canal deviations (mm) can potentially be used as another quantitative measurement of scoliosis. We then investigate how the vertebral canal is varying for different scoliosis severities ranging from normal, mild to severe C and S shape curves.



(a) Scatter plot of 2D vs 3D deviations

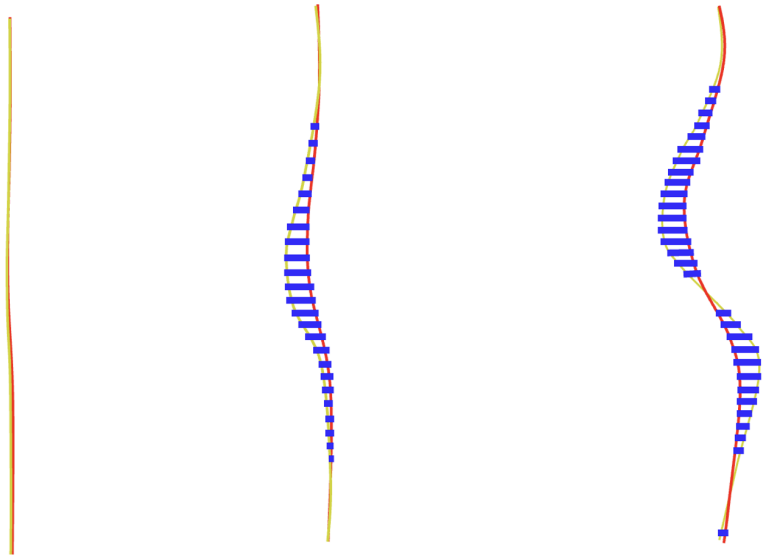


(b) Distribution of 2D deviations



(c) Zoom on mild and severe cases

Figure 4.7: **(a)** Scatter plot of spine-canal point-wise deviations (mm) from 2D coronal projection versus 3D (mm) ($\rho = 0.86, p - value < 0.05, n = 48,384$). **(b)** Histogram with density function displaying the distribution of 2D spine-canal deviation values (n = 48,384) for normal, mild and severe scoliosis cases. **(c)** Zoom in on mild and severe scoliosis cases from plot in **(b)**. The threshold for scoliosis based on human angles (greater than 6° in whole body DXA in terms of curvature is 0.083). This threshold corresponds to 2.5mm of spine-canal deviation.



(A) Normal Spine (B) Mild Scoliosis (C) Severe Scoliosis

Figure 4.8: **Visualisation of Spine and Canal Deviations for a normal (A), mild (B) and severe (C) scoliosis cases on the coronal plane.** Spine in yellow, vertebral canal in red, and deviations between the spine and vertebral canal in blue.

We can now investigate the properties of the spine that are obtained from the projections of the 3D space curve (see Figure 4.2 for a severe S shape curve).

Coronal vs Sagittal. We measured the deviation of the spine and canal at the point of maximum coronal curvature. Comparing the MRI coronal and sagittal spine-canal deviations at point of maximum coronal curvature, we observe an inverse correlation ($\rho = -0.64$, $n = 48,384$). Curves on the sagittal plane are challenging to accurately assess due to the natural variations of the spine. We notice that severe scoliosis cases tend to have straighter spines in the sagittal plane (see Figure 4.5). This inverse correlation between coronal and sagittal plane deviations is in accordance with past studies on biplanar radiographs curvature measurements [Galbusera et al. 2022]. Moreover, we observe a correspondence between the coronal plane and axial plane. The spine and vertebral canal deviation is greater on the axial projection for severe cases (see Figure 4.5).

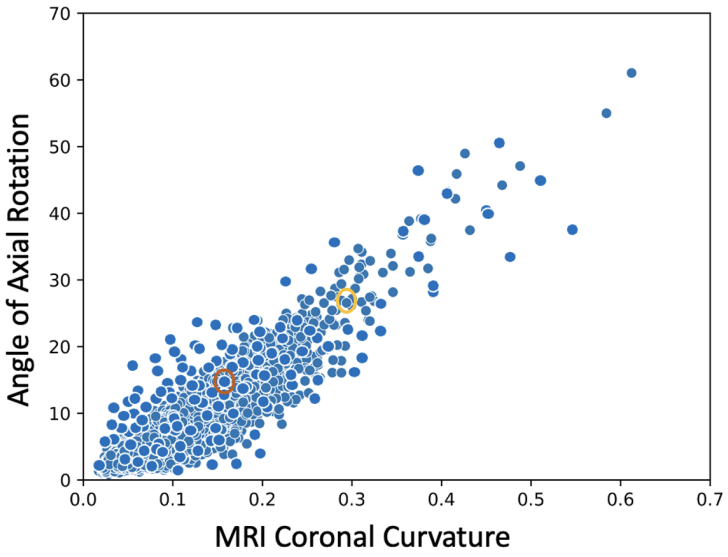
4.3.3 Curvature measurement in MRI and relation to axial plane

The correlation between maximum coronal curvature and angle of maximum axial rotation is moderately strong (ρ = 0.77, n = 48,384) which may suggest a critical role of the axial plane in relation to curvature on the coronal plane. This is in line with recent

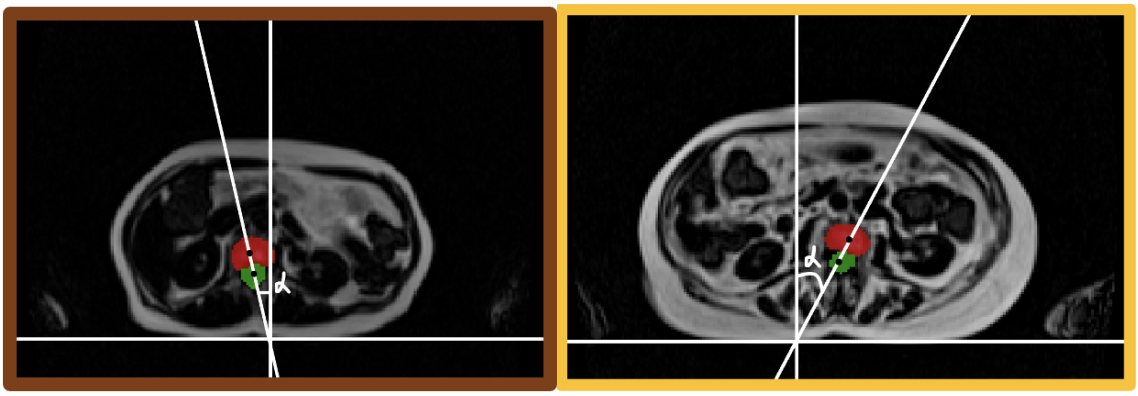
research on reconstructed 3D images [Illés et al. 2019; Karam et al. 2022]. Previous work suggested a causal link between axial deformations and onset of coronal deformations due to compensatory mechanical factors [Roaf 1958].

We show the scatter plot between the MRI axial angle of rotation at the point of maximum coronal curvature and the maximum of the MRI spine coronal projection in Figure 4.9(*a*), for all 48,384 scans in the UK Biobank. The correlation is relatively good (Pearson's ρ = 0.79) between coronal curvature and axial rotation at point of maximal curvature. This confirms the findings of Section 4.3.2, at a large scale.

As a qualitative example, we compare the spine and vertebral canal masks for a mild scoliosis case (max. abs. curvature = 0.18, brown circle in Figure 4.9(*a*)), and a more severe scoliosis case (max. abs. curvature = 0.29, yellow circle in Figure 4.9(*a*)). Axial slices corresponding to these two cases are shown in Figures 4.9(*b*) and 4.9(*c*).



(a)



(b) (c)

Figure 4.9: **(A)** Scatter plot of angle of axial rotation vs MRI coronal maximum absolute curvature (Pearson's $\rho = 0.79$, n = 48,384). Angle is given in degrees. **(B)** and **(C)** Axial slices corresponding to point of max MRI coronal curvature (yellow and brown circles in **(A)**). Spine (red) lateral deviation is more prominent on **(C)** for severe scoliosis case than **(B)** for mild scoliosis.

4.4 Conclusion

In this work, we investigated the geometry of scoliosis in 3D, while most prior work has focused on 2D deformations. We measured the curvature of the spine on one of the largest datasets of MRIs. One of the most remarkable outcomes of the visualizations is to see how the vertebral canal arranges itself to have less severe curvature than the spine itself. We also show that the axial plane is quite relevant for the assessment of scoliosis as suggested by the relatively high correlation between the angle of axial rotation and coronal curvature. By considering the spine as a 3D curve, we compared the projected 2D curves of the spine and canal on the coronal and sagittal plane. This efficient method

could be used to measure the severity of the spine's deformation.

Ultimately, the goal of this research is to provide an accurate and consistent interpretation of spinal deformations in order to support clinicians in their decision-making process. Prior to the work in this paper, the link between coronal and sagittal curves was not well defined. Also, the role of the axial plane in relation to the coronal and sagittal planes was not yet known. However, one possible future analysis could be to use the relationship between the coronal, sagittal and axial curves as a 3D classification method, without the need to explicitly model the spine in 3D, thus facilitating its adoption in clinics.

Acknowledgements

This work was supported by the Centre for Doctoral Training in Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research (SABS: R^3), University of Oxford ($EP/S024093/1$), and by the EPSRC Programme Grant Visual AI ($EP/T025872/1$). We are also grateful for the support from the Novartis-BDI Collaboration for AI in Medicine.

A Segmentation

There are four separate aligned sequences in the MRI Dixon scans used here. These are in-phase, opposed-phase, fat-only and water-only. The fat-only and water-only sequences are best suited to our task, see Figure A1. Note, the MRI scans in the UK Biobank have a lower resolution compared to typical clinical spine scans. We segment the spine using axial slices as they have higher resolution, and also support larger receptive fields for training the deep network.

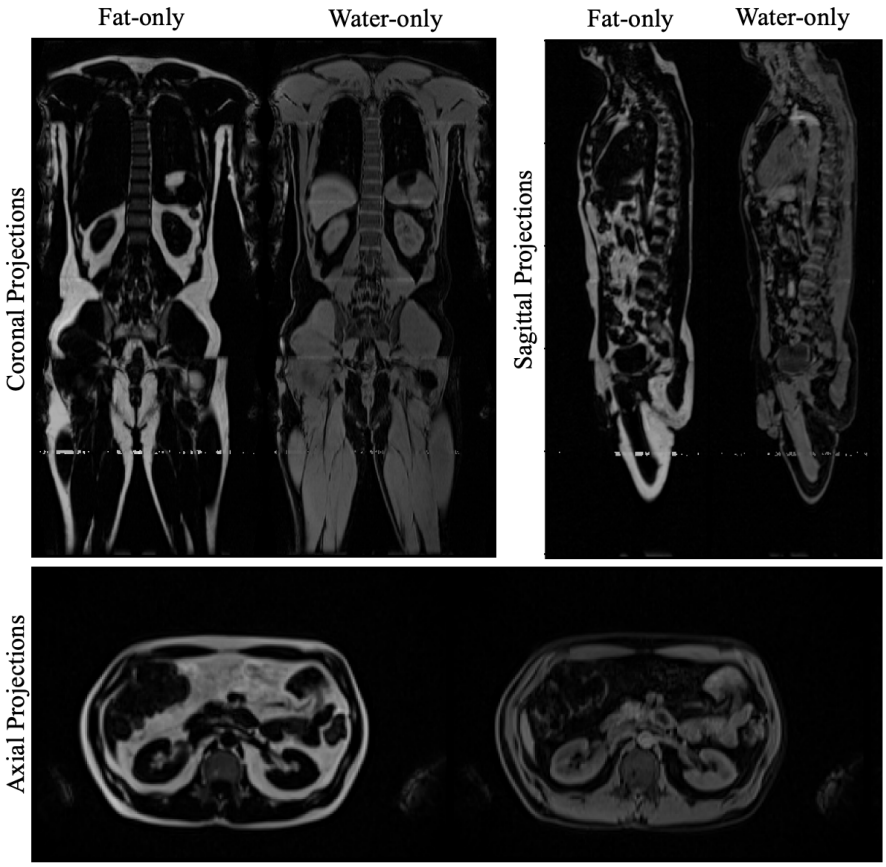


Figure A1: Coronal, sagittal and axial projections for fat-only and water-only Dixon MRI sequences.

A.1 Segmentation network

A U-Net based network architecture is used for the segmentation task [Ronneberger et al. 2015; Zhang et al. 2017]. We use a U-Net++ [Zhou et al. 2018] network with a ResNet-34 encoder. The input is $224 \times 160 \times 6$, where we stack three adjacent MRI image slices of the spine region for the two MRI sequences (fat-only and water-only). To avoid partial volume effects, and also to benefit from more context, we ingest three adjacent slices, with the middle slice as output. The output has size $224 \times 160 \times 2$, where

2 refers to the segmentation maps for the spine and vertebral canal.

For training, the loss function is a weighted sum of categorical cross-entropy loss [Yi-de et al. 2004] and Dice loss [Milletari et al. 2016] computed over a foreground/background/uncertain tri-map to mitigate potentially noisy boundaries in our labels which we define as $\pm$ 2px from the foreground boundary. Networks are trained for a maximum of 500 epochs with early stopping when the validation Dice does not increase by e^{-4}. We use self-training to leverage the whole training set i.e. n = 38,707. Inspired from the recent work on confirmation bias reduction in self-training [Chen et al. 2022], we use an independent head for pseudo-label generation to prevent potentially inaccurate pseudo-label backpropagation.

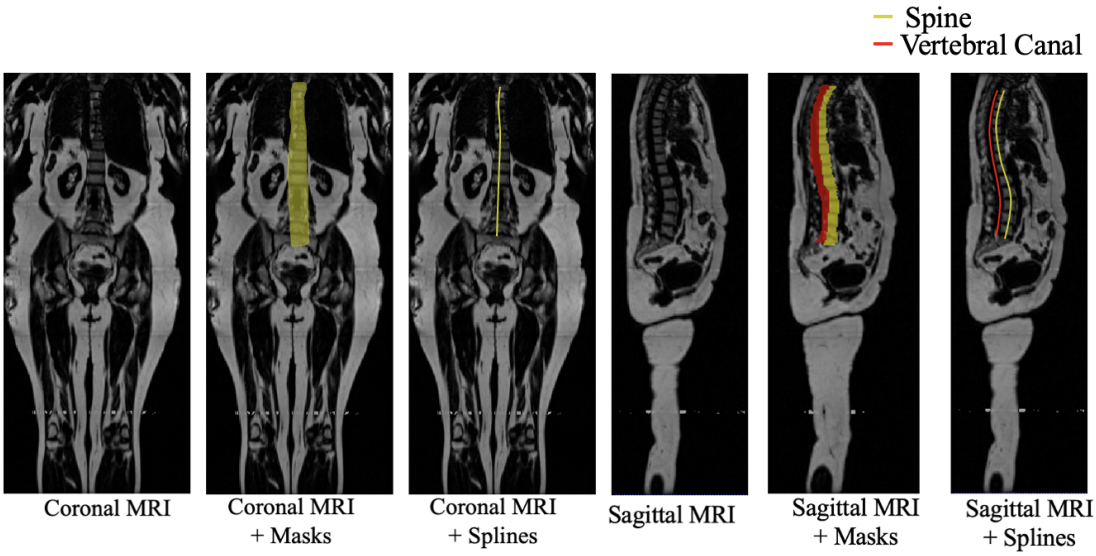


Figure A2: **Visualisation of spine and vertebral canal segmentation masks and midpoint curves on the coronal and sagittal plane.**

B Spline Fitting

B.1 2D Spline Fitting

The 2D projected points (in the coronal or sagittal planes) are approximated by a piecewise cubic spline to smooth out any noise due to sampling. For this fitting, we use the method described in [Bourigault et al. 2022].

Using a parametrised curve, we construct polynomial piecewise cubic curves. A single cubic curve has only one inflection point, but scoliosis curves may have one or more. A solution could be to add extra control points and using higher order polynomials.

However, higher order polynomials are known to be very sensitive to the locations of the control points. A common alternative in computer vision is to construct cubic curves pieced together with a greater number of inflection points. Each pair of control points form one segment of the curve, where each curve segment is a cubic with its own coefficients.

$$f_i(x) = a_i + b_i x + c_i x^2 + d_i x^3 \tag{4.3}$$

where f is the function representing the curve between control points i and i + 1.

We ensure C^0, C^1, C^2 continuity conditions.

- C^0: Each segment is required to pass through its control points. That is, $f_i(x_i)$ = y_i, and $f_i(x_{i+1})$ = y_{i+1}
- C^1: Each curve segment has the same slope at each junction, $f_i^{'}(x_{i+1})$ = $f_{i+1}^{'}(x_{i+1})$
- Each curve segment has the same curvature at each junction, $f_i^{''}(x_{i+1})$ = $f_{i+1}^{''}(x_{i+1})$

We improve the method in [Bourigault et al. 2022] by changing the uniform placement of a fixed number of knots by automatic knot selection using penalised regression splines [Ruppert et al. 2003]. The spline curve is composed of $n-1$ piecewise cubic polynomials where n is the total number of knots. The number of knots is selected in the range from 2 to 10.

n is optimised using a penalty to balance goodness-of-fit and smoothness. The selection of knots is such that the model chooses from a bigger selection of functions. As the number of knots increases, the model overfits the data. Too few knots on the other hand gives a more restrictive function.

B.2 3D Spline Fitting

We now extend the 2D spline fitting to three-dimensional space. We have two systems of linear equations for x and y: $M_x\mathbf{b}_x$ = **x** and $M_y\mathbf{b}_y$ = **y**, where **b** is the vector of curve coefficients, **y** is the vector of constants, and M is a matrix of continuity conditions ie. C^0, C^1, and C^2. Each system is solved similarly as in 2D section above, except that we are solving two linear systems instead of one.

Chapter 5

3D Spine Shape from 2D DXA

This paper was published in the proceedings of the Medical Image Computing and Computer Assisted Intervention Conference (MICCAI), 2024 [Bourigault et al. 2024b].

The analysis presented here builds upon the foundational work on 2D scoliosis assessment detailed in the previous Chapter 3 and 3D scoliosis in Chapter 4. While earlier work focused on the limitations of the Cobb angle and the initial efforts to quantify scoliosis from two-dimensional imaging, this study aims at regressing the 3D spine shape from a 2D DXA scan. We addressed modality discrepancies through a two-stage alignment process and we leveraged advanced deep learning techniques (e.g., transformer-based regression) to improve prediction accuracy and robustness. We showed that our transformer-based regression model is effective at integrating DXA and MRI information to provide precise 3D representations of the spine.

This evolution from a 2D-centric approach to one that embraces three-dimensional reconstruction at the patient-level not only deepens our understanding of spinal morphology but also lays the groundwork for future diagnostic and screening tools in clinical practice.

3D Spine Shape from 2D DXA

Emmanuelle Bourigault Amir Jamaludin

Andrew Zisserman

VGG, Department of Engineering Science, University of Oxford

emmanuelle,amirj,az@robots.ox.ac.uk

Abstract

Most recent work on 3D analysis of scoliosis have been performed on EOS. The resolution of EOS is higher than other modalities and these biplanar images are acquired simultaneously which enables accurate 3D spine reconstruction [McKenna et al. 2012; Carreau et al. 2014]. However this modality is costly and not widely available. In this work we make use of a more affordable approach using DXA scans which are low radiation X-rays and increasingly being used in the screening of scoliosis. Several work have been proposed to segment and reconstruct the 3D spine with heuristics. These methods mainly rely on atlas or shape-based models for spine reconstruction. In contrast, in this work we do not use shape-based models and instead learn two regress two orthogonal projections of the spine from one AP DXA scan. Via minor post-processing, we showed that we could reconstruct the whole 3D shape of the spine.

2D-3D, Symbiosis, Scoliosis, MRI, DXA

5.1 Introduction

The standard procedure to examine the spine for the presence of scoliosis is using antero-posterior (AP) X-rays and measuring the angle between the most tilted vertebrae [Cobb 1948]. Scoliosis typically affects growing children and proper diagnosis of scoliosis requires multiple follow-up scans. As such, Dual-energy X-ray (DXA) scans, with its lower radiation dose than X-rays is quickly becoming an acceptable alternative [Taylor

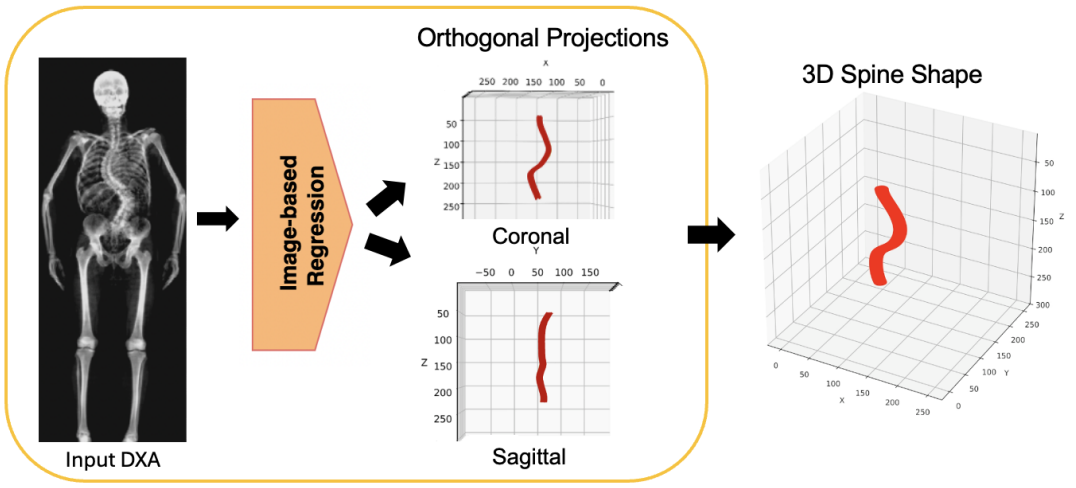


Figure 5.1: **Inference.** Given a DXA scan, the model predicts the coronal and sagittal projections of the 3D spine. Once these two orthogonal views of the spine are obtained, the 3D spine can be reconstructed. A visualization of the rotating spine is available at: https://www.robots.ox.ac.uk/~vgg/research/dxa-to-3d.

et al. 2013; Jamaludin et al. 2018; Ng et al. 2023; Jamaludin et al. 2019b]. However, this still does not solve the fundamental issue of scoliosis diagnosis; it is essentially a 3D disorder but the focus is on the 2D lateral shift of the spine. Imaging this disorder on a Magnetic Resonance Imaging (MRI) is a good alternative, and multiple studies have shown that there is indeed more useful information that can be extracted from MRIs for scoliosis [Illés et al. 2010; Donzelli et al. 2015; Bourigault et al. 2023]. MRI however is more expensive and requires more time for one single scan compared to X-ray/DXA.

In this paper, our objective is to obtain the 3D patient-specific spine shape from a 2D DXA. We explore whether this is even possible given that 3D information is "lost" in the projection of a 2D DXA scan. We show that it is possible to learn to infer this "lost" information by leveraging a paired imaging set of DXA scans with corresponding MRIs taken at roughly the same time. The MRI provides the 3D geometry of the spine, and a network can be trained to map the 3D spine from the 2D DXA. We achieve this task by regressing coronal and sagittal curve projections. Through these two orthogonal curves, coronal and sagittal, we can recover the 3D geometry of the spine.

5.1.1 Related Work

To date, the vast majority of scoliosis research has focused on 2D shape analysis of the spine or part of the 3D spine [Taylor et al. 2013; Jamaludin et al. 2016; Jamaludin et al. 2018; Windsor et al. 2020; Bourigault et al. 2022; Ng et al. 2023; Jamaludin et al. 2019b].

Limitations of 2D spine analysis arise particularly in classifying 3D curve shape. Indeed, deviations are not limited to the coronal plane, they include twisting of the spine in multiple directions [Rockenfeller and Müller 2022], and the importance of other planes, e.g. for axial rotations, is well recognized [Illés et al. 2019]. Recently, there has been a growing interest in the community on EOS imaging with simultaneous acquisition of coronal and sagittal views producing 3D spine shape [Rehm et al. 2017]. This enables, for example, a better evaluation of the effect of brace surgical treatment [Courvoisier et al. 2014; Dubousset et al. 2014].

There are several works with similar goal to ours; the most similar work is by [López Picazo et al. 2018] which works on DXA scans and predicts the 3D model of the spine using statistical shape models (SSM) [Cootes et al. 1995]. Other works use biplanar X-rays either with SSM [Aubert et al. 2019; Benameur et al. 2003; Clogenson et al. 2015] or contour matching [Zhang et al. 2013]. In our work, we directly estimate the 3D spine from a single 2D DXA. Our 3D regression involves only a single pass through a feed-forward network.

5.2 3D From 2D

Our method for estimating 3D spine geometry from a single 2D DXA is simple. We essentially learn to regress 2D curves; first by directly regressing the curve of the spine extracted from the 2D DXA itself (this is a coronal projection of the 3D spine), and then by predicting the sagittal projection of the 3D spine curve. Given these two orthogonal projections (coronal $x(z)$ and sagittal $y(z)$), the 3D curve can trivially be obtained (as $(x(z), y(z))$). Learning the sagittal projection is only feasible through the use of a large-scale public dataset, consisting of *paired* whole-body DXA and MRI of the same subjects [Sudlow et al. 2015], where the 3D spine curve can be extracted from the MRI. The challenges of this problem are: (i) alignment of the paired DXA and MRIs, discussed in Section 5.2.2, and (ii) how to directly regress 2D curves from the DXA scan, discussed in Section 5.2.3.

5.2.1 Problem definition

The problem consists of regressing a 3D spine curve from a DXA image. This can be separated into two separate regression problems, namely: (i) the regression of the coronal

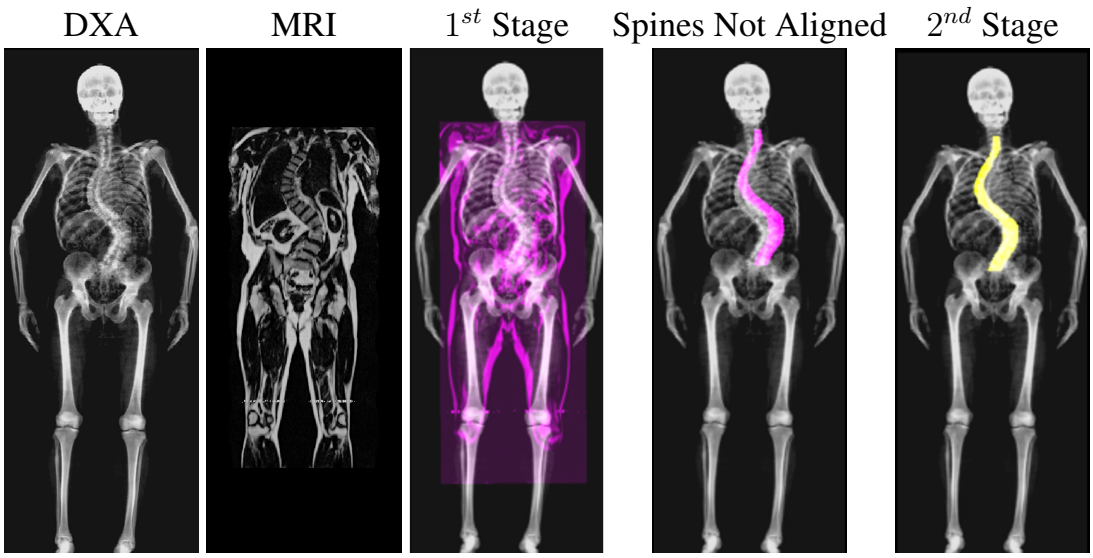


Figure 5.2: **DXA to MRI two-Stage Alignment**. The two scans are iteratively aligned using a three parameter planar transformation. From left to right: DXA scan; original coronal projection of MRI scan (not-aligned to the DXA); overlay of MRI aligned to DXA after the image-level alignment first stage; overlay of segmented spines after the first stage; overlay of spine segmentation after the spine-level alignment second stage.

or AP curve, and (ii) the regression of the sagittal or lateral curve. For each of the 2D projections, we define three sets of points $P_i = \{(x_i(z), y_i(z)), z \in| \ 1 \leq z \leq 209 \}, i = 1, 2, 3$ for the central and two lateral curves of the spine respectively, and z is the vertical height of the scans (normalised between 1 and 209). These define the segmentation and mid-curve of the spine in the 2D projections. Our objective is to regress these curves (in the coronal and sagittal planes) from the DXA scan.

5.2.2 Modality Alignment

The paired images, the 2D DXA scans and the 3D MRIs, do not come registered. Hence, the first step we take is to align these two modalities. As we are interested in inferring the 3D information from 2D images; we register the 3D MRI and the extracted spine curve to the 2D DXA. The alignment proceeds iteratively in two stages: (i) a rough image-to-image alignment of the two imaging modalities followed by, (ii) a finer alignment of the extracted segmentation/curve of the spine from the MRI to the segmentation/curve of the spine from the DXA. See Figure 5.2.

For the rough alignment, we use a pipeline proposed by [Windsor et al. 2021] which finds the best 2D transformation to align the 3D MRI (via its coronal projection) to the 2D DXA. The 2D transformations involve three parameters (a rotation angle θ, and two

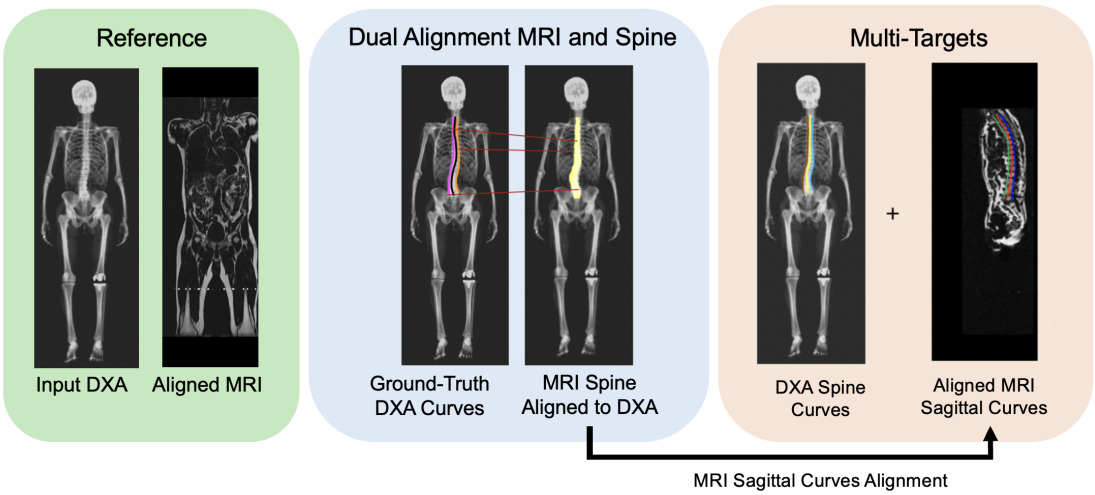


Figure 5.3: **Model Learning**. The regression model is learnt from pairs of aligned DXA and MRI scans. The regression targets are the DXA curve, and the sagittal curve (projected from the 3D MRI spine). The alignment for the sagittal curve to DXA is obtained from the alignment of the coronal projection of the 3D MRI to the DXA. Six curves are regressed: the centerline of the spine as well as the left and right boundaries of the segmentation, for both the coronal and sagittal views.

translations). To compute the transformation, 10 rotation angles are sampled in the range $[-2, 2]$ degrees, and the translation is obtained by convolving CNN spatial-feature maps of the DXA and MRI, and selecting the point of maximum response.

For the second stage, we align the spine curves between the two modalities. Again, the 2D transformation consists of a rotation and translation. We use keypoint matching sampled along the spine contour of the DXA, and compute the transformation that minimises the mean squared error (MSE).

It is possible that the person changed position too much between the DXA and MRI scans, and it is not possible to align the spines due to deformations. To check for this we measure the overlap of the spine segmented in the DXA scan with the projection of the 3D spine segmented in the MRI (see Figure 5.2). We apply a threshold for filtering out the poorly aligned scans: if the Intersection-over-Union (IoU) of spine masks is below 70%, then they are discarded. A total of 14,065 paired DXA-MRIs are rejected by this test which is 28.8% of the data.

5.2.3 Learning the Regressor

In total, we regress 6 curves using a single model. The target curves for the regression are obtained from the DXA for the coronal view, and from the projection of the 3D MRI for the sagittal view (see Figure 5.3).

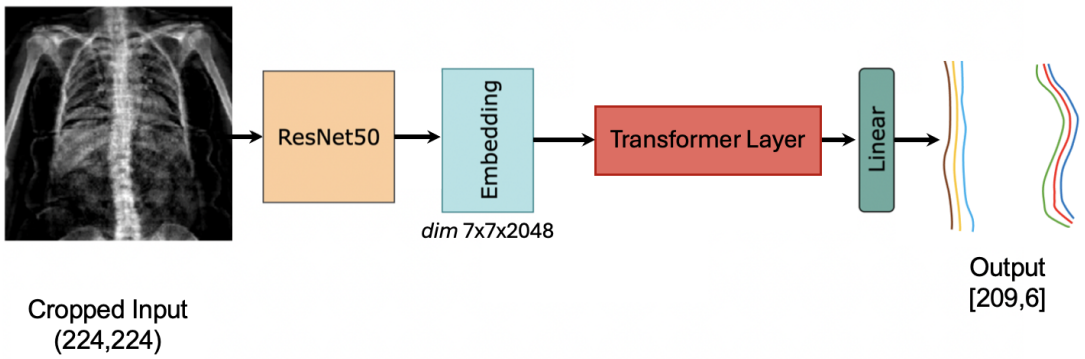


Figure 5.4: **Image-Based Regression of Coronal and Sagittal Spine Curves.** We use a ResNet50, pre-trained on ImageNet-21k, with a transformer layer to regress the spine curves $(x_{(1,2,3)}(z), y_{(1,2,3)}(z)), z \in [1, 209]$ for left, center and right curves. The feature map extracted from ResNet50 are of resolution 7 x 7 x 2048, each vector feature from ResNet50 (49 x 2048) is used as input into a transformer layer. The model regresses the 6 curves (209 x 6) where we have 6 vectors for the 6 output spine curves, of dimension 209. Detailed Architecture in Appendix A and Figure A1.

Encoder. Our regressor consists of a multi-scale feature extractor coupled with a transformer layer to learn the long-range dependencies more effectively (see Figure 5.4). The initial part of the architecture consists of an image feature extractor with ResNet50 pre-trained on ImageNet-21k. The feature map of the penultimate convolutional block is extracted as $7 \times 7 \times 2048$ and fed into a standard transformer layer. The output of the transformer layer is then average pooled before the ultimate linear layer. The regression head consists of a linear layer to predict the output curve points from the transformer output vector.

Loss. Our loss is the L_1 difference between the target and predicted spine curve points: $L = \sum_{i=1}^{n} |x^i - \widehat{x}^i|$, where n is the number of samples, x^i are the ground-truth spine points, and $\widehat{x^i}$ the predicted spine points at a given z^i.

We regress three curves for each view (coronal (x) and sagittal (y)): the central curve of the spine as well as the left and right curves bounding the segmentation. The central and lateral curve points are (x_1,x_2,x_3,y_1,y_2,y_3), where 1 and 3 are either the right/left or anterior/posterior curves depending on the plane projection, and 2 is the central curve of the spine.

5.3 Dataset & Implementation Details

The UKBiobank is a publicly available dataset of 48,384 full-body Dixon MRI paired with DXA scans [Sudlow et al. 2015]. The UK Biobank MRIs are resampled to be isotropic and cropped to a consistent resolution ($700 \times 224 \times 224$). The dataset is split into 80:10:10 for training (27,816), validation (3,477), and testing (3,477) after filtering the non-aligned scans by the dual alignment procedure explained in Section 5.2.2. Our training-validation-test is balanced in scoliosis cases, each containing 20% of scoliosis cases. Our model takes as input a cropped DXA image of the whole spine. This is achieved using a 224×224 cropping window of the spine given spine segmentation from [Bourigault et al. 2022]. This is done by computing the midpoint of the segment joining the endpoints of the spine from the segmentation mask and using it as the centre of the square window. The original DXA resolution is (832×320).

Obtaining the 2D spine from the DXA images. We obtain the spine centroids and 2D spine mask segmentation in DXA which involves spine segmentation using pseudo-labelling in an active learning framework [Jamaludin et al. 2018; Bourigault et al. 2022]. This is used as the target for the coronal view regressor.

Obtaining the 3D spine from the MRIs. We used a segmentation network to obtain the 3D whole spine curve in the MRI (3D centroids and segmentation) [Bourigault et al. 2023] trained on adjacent axial slices (n-1,n,n+1) to limit the loss of depth information. These labels are used for training our sagittal regression model.

5.3.1 Implementation Details

The regressor uses a the Bottleneck Transformer [Srinivas et al. 2021] together with a ResNet-50 image encoder. Details on the model architecture are given in the Appendix (see Section A).

For curve regression, all DXA inputs and target spine curves are normalised to fixed height for spine ranging from 1 to 209 pixels. We train our model for 500 epochs. We use five$-fold$ cross-validation, where we repeat validation on 5 stratified folds. The batch size is set to 16, optimizer is Adam [Kingma and Ba 2015] with $\beta = (0.9, 0.999)$, and the learning rate is initially set to $1e^{-4}$ with decay every 200 epochs. We used one 32GB Tesla V100 GPU. To reduce overfitting, we employ two techniques, using dropout with a probability of $p = 0.3$ and we employ a regularizer to the L1 loss. We also use

different augmentation techniques with cropping, image contrast and random Gaussian noise in training.

5.4 Results

5.4.1 Evaluation Metrics

We measure the performance of our model using the mean absolute error $MAE = \frac{1}{n}\sum_{z=1}^{n}|y_z - \widehat{y_z}|$, and the mean of the relative error $RE = \frac{1}{n}\sum_{z=1}^{n}\frac{|y_z - \widehat{y_z}|}{\widehat{y_z}}$ between the predicted points $\widehat{y_z}$ and the ground truth $y_z, z \in [1, 209]$. The mean of RE is always between 0 and 1, the lower the relative error the better. To assess the accuracy of whole spine predicted masks obtained from the area between the lateral curves compared to project ground-truth masks obtained from segmentation of the MRI, we use the 2D Intersection-over-Union, IoU. We compute the 3D IoU to evaluate 3D spine shape estimation from our model given reference labels from MRIs, as explained in Section B.

5.4.2 Spine Curves Estimation and Robustness

The DXA and Coronal MRI spine curves are predicted at sub-pixel level precision with better precision and lower variance for the combined ResNet50+Transformer framework ($MAE = 0.66 \pm 0.21$, $RE = 0.084 \pm 0.02$, $IoU = 89.6$) compared to ResNet50 or ViT (see Table 5.1). We show that using residual blocks with a transformer layer, our model can reliably estimate not only the coronal MRI projection from a single DXA but any spine curve projections along a 360° rotation about the vertical axis such as the sagittal projections ($MAE = 1.65 \pm 0.6$, $RE = 0.21 \pm 0.08$, $IoU = 86.8$), see Table 5.1). Our model is also able to capture the pattern of curves with straighter sagittal curves for more severe cases of scoliosis (see Figure 5.5).

3D evaluation. We also evaluate the performance of our model to reconstruct 3D spines from 2D DXAs. We effectively have a segmentation of the spine bounded by the lateral curves from our pipeline. To obtain the 3D spine masks from two 2D segmentation, we use the 4 points in antero-posterior and left-right to generate a series of bounding axial plane ellipses as we go down in z (see Appendix B). We use an untouched set of 150 MRIs from the UKBiobank manually annotated spines from [Bourigault et al. 2023] to measure the IoU. The IoU between the predicted and ground-truth 3D spine masks,

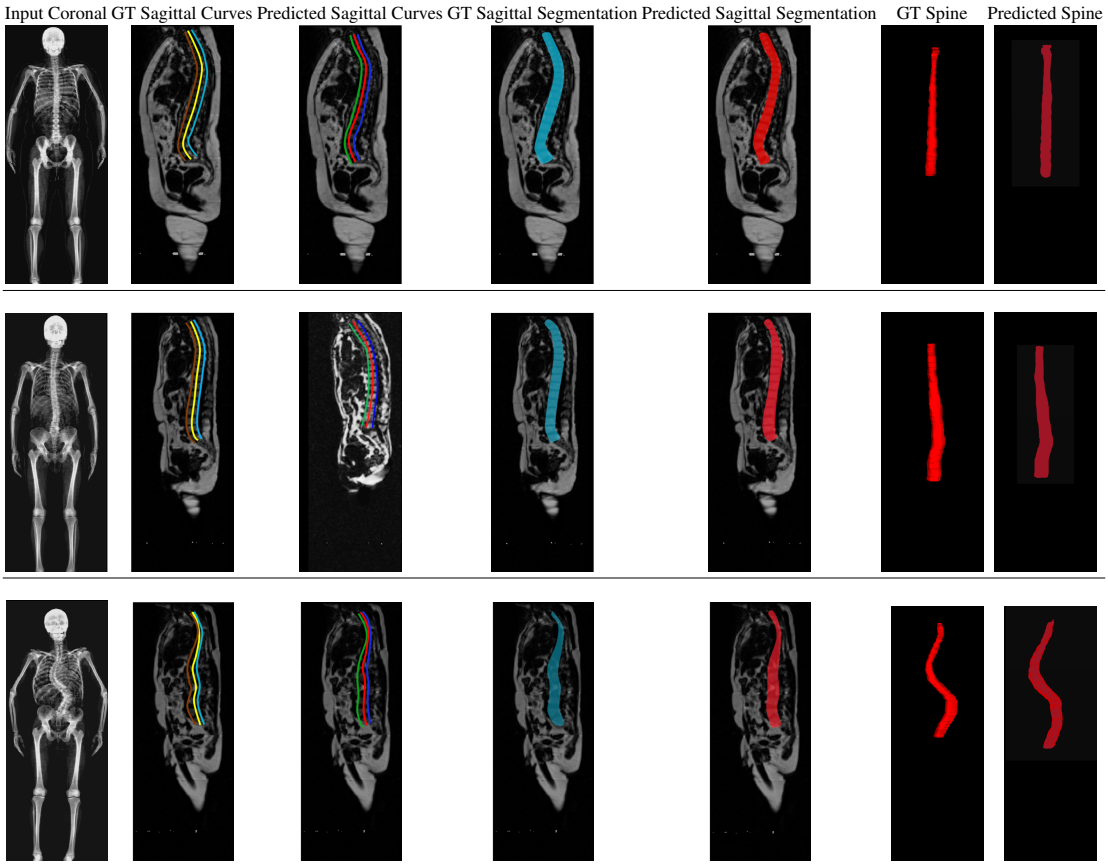


Figure 5.5: **Qualitative Results of DXA to MRI Sagittal Curve Generation on the Test Set.** We show from left to right, the input DXA, the ground-truth sagittal curves, predicted sagittal curves, ground-truth spine mask, predicted spine mask overlayed on MRI sagittal slice (n=112), the ground-truth 3D spine, and its prediction using Gaussian rendering. The severity of the scoliosis increases from top to bottom.

averaged over the test set, is 83.8 ± 1.1. We also measure the IoU for detected spine. See Figure 5.6 for visualising the mean average precision at different IoU thresholds. We also measure the average deviation of the spine curve point-wise along the spine as a metric for 3D error. Our model achieves a 3D average spine curve deviation of 1.42 voxels or $3.12mm$.

Discussion. One probable reason that the model is able to infer a sagittal view from coronal DXA is that it has access to the entire (cropped) DXA scan, and there are at least two cues it can use: (i) the different intensities of the imaged bones of the spine give information about their angle and depth; and (ii) the position of the rib cage in the image depends on the spine, and so indirectly the 2D layout of the ribs in the image gives information about the spine.

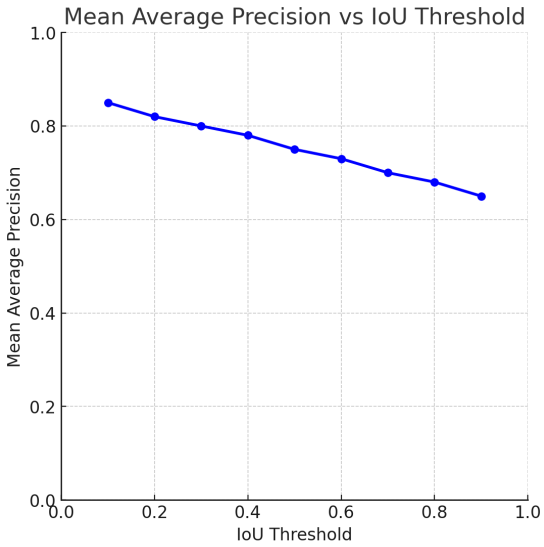


Figure 5.6: **Mean Average Precision of Predicted Spine Masks for Different IoU Thresholds over the Test samples.** We measure the performance of the 3D segmentation using 3D IoU. IoU thresholds ranges from 0.1 to 0.9, with steps of 0.1.

Target	Model	Spine Curves Absolute Error			Spine Curves Relative Error			Spine Mask (IoU)
		Mean	Median	SD	Mean	Median	SD	
DXA	ResNet50	0.71	0.70	± 0.26	0.08	0.076	± 0.03	91.4
	ViT	0.69	0.67	± 0.24	0.073	0.071	± 0.02	91.9
	ResNet50 + Transformer	0.58	0.57	± 0.18	0.062	0.057	± 0.01	92.3
Coronal MRI	ResNet50	0.81	0.79	± 0.3	0.11	0.094	± 0.05	88.3
	ViT	0.78	0.78	± 0.28	0.091	0.083	± 0.03	88.9
	ResNet50 + Transformer	0.66	0.64	± 0.21	0.084	0.079	± 0.02	89.6
Sagittal MRI	ResNet50	3.63	3.29	± 1.2	0.41	0.39	± 0.14	83.6
	ViT	2.99	2.73	± 1.1	0.38	0.36	± 0.11	84.1
	ResNet50 + Transformer	1.65	1.58	± 0.6	0.26	0.21	± 0.08	86.8

Table 5.1: **Performance of Image-based Models for Spine Curves Regression.** For a given input DXA, we predict curves from the DXA target (*1st block*) or curves from the corresponding MRIs (*2nd and 3rd blocks*). We predict 3 curves (center + 2 laterals) for each block. The *2nd and 3rd blocks* represent a single multi-view model outputting coronal and sagittal curves. The *Target* column specifies the model configuration with output curve modality (DXA for baseline, coronal and sagittal MRI). The *Model* column shows the different models. Then, we present the *Absolute Error* and *Relative Error* in terms of mean, median, and standard deviation of spine curves predicted versus reference curves (in pixels). The far-right column shows the 2D IoU of the masks bounded by the lateral spine curves.

5.5 Conclusion

Our model is able to give a patient-specific representation of the spine in 3D from a single DXA scan. We show that our method is effective in capturing the intricacies of the 3D spine. As such, this work has the potential to assist in the diagnosis and screening of scoliosis and other spinal disorders. Our primary focus for future work includes investigating confidence prediction for the sagittal curve.

Acknowledgements

This work was supported by the Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research (SABS: R^3) Centre for Doctoral Training ($EP/S024093/1$), the EPSRC Programme Grant Visual AI ($EP/T025872/1$), and the Novartis-BDI Collaboration. This work has been conducted using the UK Biobank resource (application number 17295).

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

A Implementation Details and Ablation

Regression Network of Spine Curves. In this work we adopt a lightweight transformer layer on top of ResNet50 (see Figure A1). Before the final average pooling, we take the 7 x 7 x 2048 ResNet50 feature maps and flatten them to obtain 49 embedding tokens. These vectors are the input to the transformer layer. We follow the BotNet architecture [Srinivas et al. 2021] for the relative positioning encoding in the transformer layer. The output of the transformer layer is then average pooled before the ultimate linear layer.

Ablation Experiments. We experiment with varying input sample size during training with an overall improved test performance for spine mask prediction of +3.4 IoU using the whole training set of 30k available versus 500 samples. Therefore, the network benefits from training on large datasets and it improves its ability to generalise. We also experiment adding more than one transformer layer on top of ResNet50. This does not significantly boost performance on the order of +0.002px and +0.003px average improvement for coronal and sagittal curve regression respectively.

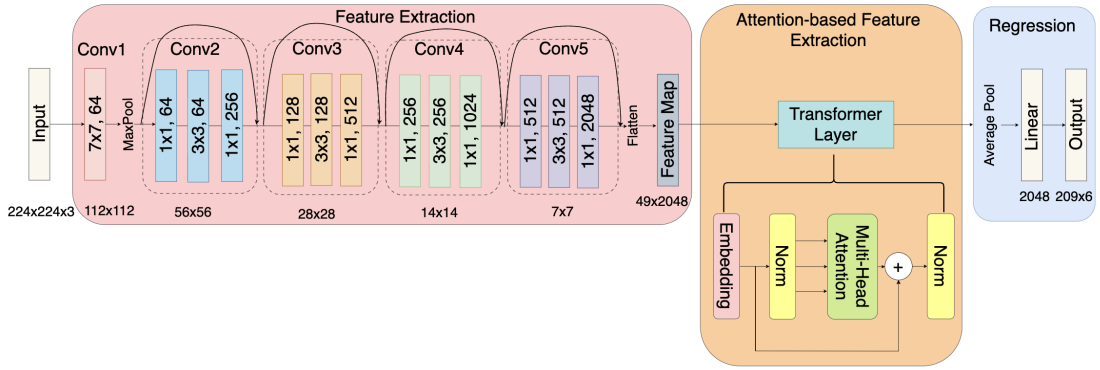


Figure A1: **Full Architecture of our ResNet50 with Transformer Layer.**

B 3D Spine from 2D Projections

Computing 3D Spine Shape From 2D Orthogonal Curves. In this section, we outline the steps taken for 3D spine shape recovery from two 2D planes i.e. coronal and sagittal. The output of our network are 2D spine curves on the coronal (XY) plane and sagittal (YZ) plane. We are able to reconstruct the 3D spine shape with minor post-processing. The idea is simple, fitting ellipses in the axial plane along the spine from top to bottom (see Figure B2). We ensure the ellipses go through predicted points from the two lateral coronal (right and left) and sagittal (antero and posterior) curves.

More examples of predictions of 3D spine shape from 2D DXA are available on our website https://www.robots.ox.ac.uk/~vgg/research/dxa-to-3d. Our model works well in estimating sagittal MRI projections for normal spines and severe scoliosis spines.

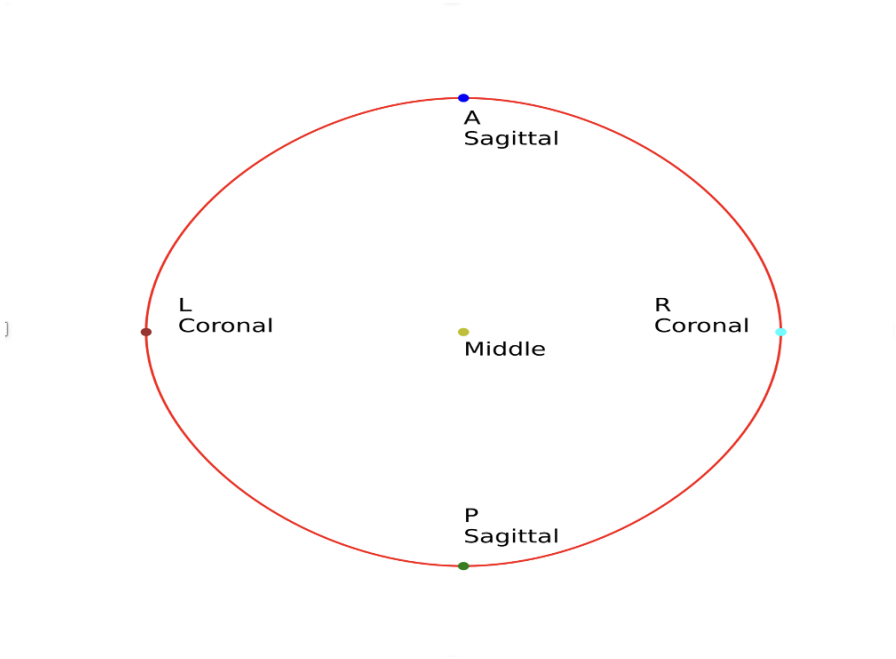


Figure B2: **3D Spine Shape Reconstruction through Ellipse Fitting.** We form ellipses in the axial plane along the spine top to bottom that go through key points from the orthogonal curves obtained by our model. We make sure the ellipses go through the antero-posterior (A-P) sagittal curves and the right-left (R-L) coronal curves. Each dot is a point on the 6 curves shown on Figure 5.4. The middle point of the ellipse is made from the aligned mid coronal and mid sagittal curves.

Chapter 6

Automated DXA Scoliosis Method

The paper is under submission for European Spine Journal, 2025.

In this study, we developped and validated an automated method for maximum angle measurement and prediction of curve pattern on a reduced set of DXA scans from the UK Biobank that have been annotated by expert clinicians.

In Chapter 3, we showed we could reliably measure scoliosis on a continuous scale by extracting the spine curve from the segmentation of the spine and measure its curvature. However, curvature is not how clinicians measure scoliosis in practice. The gold-standard is the Cobb angle, implying that novel approaches would need to be compared to Cobb angles. Furthermore, there is not a one-to-one mapping between curvature and Cobb angles. In the optic of matching more closely human annotation, we proceed by mimicking the way human annotated DXA scans. We propose in this work a fully automated way of measuring the human maximum angle as defined by [Taylor et al. 2013] on a reduced annotated set of 1,929 scans from the UK Biobank, of which 308 have been annotated with angles. We validate our method for automatically measuring human maximum angle against a follow-up session taken a year apart. We show we can also measure curve patterns (direction, location, and number of curves) accurately.

Automated DXA Scoliosis Method

Emmanuelle Bourigault
VGG, University of Oxford

Amir Jamaludin
VGG, University of Oxford

Emma M. Clark
University of Bristol

Jeremy Fairbank
Nuffield Department of Orthopaedics, University of Oxford

Timor Kadir
Plexalis Ltd

Andrew Zisserman
VGG, University of Oxford

emmanuelle,amirj,az@robots.ox.ac.uk

Abstract

This paper is about automating and validating a total body dual energy X-ray absorptiometry (DXA) scoliosis method (DSM) on a sample of 1,929 UK Biobank annotated scans of which 308 have scoliosis and have been annotated for angles. We automatically measure the angle, the direction, and the location of the apex on the 308 scans that have been annotated for maximum angle of the spine. The method can be broken down into two main steps: (i) compute the mid-curve of the spine from the DXA, and (ii) analyse the geometry of the curve to find the maximum angle of the spine which represents its severity. To validate our method, we manually annotated a proportion of UK Biobank DXA scans (n=1,929) with

angle measurements of the largest curve, the direction of the curve, and the number of curves. There is very good agreement between the manual and predicted angle of the spine with Pearson's correlation value of 0.89, and a mean difference of less than 5 degrees between manual and automated readings with 95% confidence. Our model accurately predicted curve location (specificity of 0.883), direction (specificity of 0.779) and number of curves (specificity of 0.671). There has long been uncertainty around the role of screening for scoliosis. We expect these studies to enable accurate annotation of scoliosis in very large datasets that include spinal imaging (commonly DEXA and MRI) for definition of scoliosis phenotypes and genotypes in epidemiology, screening and clinical applications.

Deep learning, Scoliosis, Geometry, Epidemiology

6.1 Introduction

Scoliosis is a spinal deformity that, if left untreated, may require invasive surgery and lead to long-term back pain. The UK Biobank dataset used in this study consists of adults aged 40–80 years. Most automated methods for scoliosis measurement focus on Adolescent Idiopathic Scoliosis (AIS), but in this work, the aim is to measure scoliosis in adults using DXA scans from the UK Biobank. While AIS is a primary concern in adolescents, adults can develop degenerative scoliosis, which results from the wear and tear on spinal discs over time. In AIS, right thoracic curves are predominant [Konieczny et al. 2013], but there is limited research on the shape of the spine in adult scoliosis. Investigating the curvature patterns in both adolescents and adults could lead to better classification and diagnosis of scoliosis. Degenerative scoliosis, common in the aging population, is associated with disc degeneration and vertebral body collapse. The goal of this research is to validate a new, fully automated method for quantifying the size of the spinal curve from DXA scans using machine learning techniques, applied for the first time to an adult cohort from the UK Biobank. This work builds upon methods developed in the SpineNet software [Jamaludin et al. 2016]. Recent advancements in deep learning have significantly enhanced the automation of Cobb angle measurements from X-ray images, a critical factor in diagnosing and monitoring scoliosis. Several studies have explored various methodologies to improve the accuracy and reliability of these automated measurements [Li et al. 2024a; Shao et al. 2024; Zhu et al. 2024]. We propose a geometric approach to evaluate spinal deformities. Our results demonstrate

that automated spine angle measurements, on a continuous scale, align closely with human measurements from an annotated dataset. This method could improve scoliosis diagnosis by providing a more comprehensive set of quantitative information, including curve location, direction, and extent. The overarching goal of this study is to define a scoliosis phenotype that facilitates the identification of relevant predictors for curve progression. By utilizing population-based cohorts, we aim to establish relationships between automated spine angle estimations and other potential biomarkers of scoliosis, such as body composition, age, and anthropometric measurements [Clark et al. 2014].

6.2 Related Work/Background

6.2.1 Scoliosis Research

Progression of Scoliosis. Scoliosis is a complex three-dimensional spinal deformity with variable progression patterns across different life stages. While idiopathic scoliosis (IS) often stabilizes after skeletal maturity [Weinstein et al. 2019], certain risk factor such as growth velocity, curve magnitude at diagnosis, and skeletal immaturity can influence progression [Lonstein and Carlson 1984; Sanders et al. 2008]. Longitudinal studies suggest that adolescent idiopathic scoliosis (AIS) progresses most rapidly during peak pubertal growth [Weinstein et al. 2019; Clark et al. 2014]. However, adult scoliosis may also worsen due to degenerative changes, particularly in curves exceeding 30° [Marty-Poumarat et al. 2007]. Recent work by [Cheung et al. 2022] highlights biomechanical and genetic factors (e.g., LBX1 and GPR126 gene variants) that may predispose individuals to progression, though predictive models remain underdeveloped.

Measurement Techniques and Reliability Challenges. The Cobb angle, introduced by [Cobb 1948], remains the gold standard for quantifying scoliosis severity. Despite its widespread adoption, studies report inter-observer variability of 3° to 10° and intra-observer variability of 2° to 7°, limiting reproducibility [Morrissy et al. 1990; Carman et al. 1989]. Efforts to standardize measurements include digital tools [Komeili et al. 2019], yet manual assessments still dominate clinical practice. In addition, repeated radiographs pose long-term risks, necessitating low-dose alternatives i.e. EOS imaging [Stokes et al. 2018], which is not widely adopted due to the high cost of procedures. Affordable alternative modalities like dual-energy X-ray absorptiometry (DXA) offer a lower radiation exposure but may underestimate curves due to supine positioning (Clark et al. (2014)).

A 6° threshold is recommended for DXA-based diagnosis [Clark et al. 2014], though debate persists about its sensitivity compared to standing radiographs [Stokes et al. 2018].

6.2.2 DSM Method

The DSM [Taylor et al. 2013] uses a modified-Ferguson method, by drawing a "normal spine line" (NSL) through the centre of the spine level with the first rib attachment, down to the centre of the spine at L5. Then, the apex of the curve is identified. Lines are drawn from the apex of the curve to the NSL at the point where the centre of the spinal column first touches the NSL on return from the apex (see Figure 6.1).

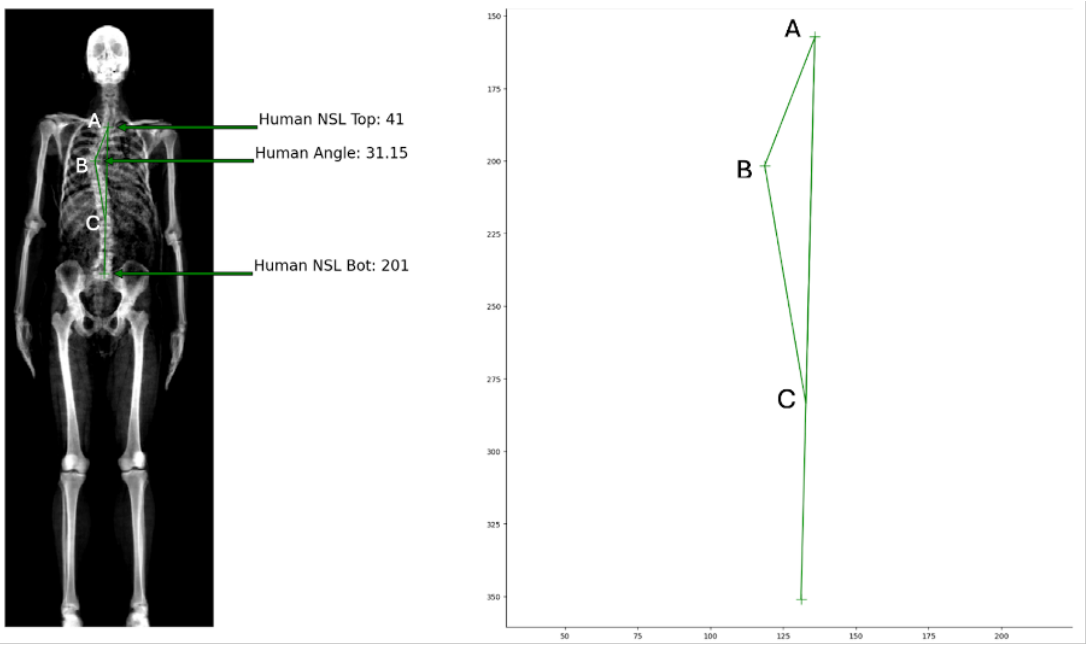


Figure 6.1: **DXA Scoliosis Method (DSM).** We show how the maximum angle of the spine is measured by humans. First the normal spine line (NSL) is manually drawn from the region of the clavicle (Human Top NSL) to the last lumbar of the spine (Human Bottom NSL). Then, the apex of the curve (green cross) corresponding to the maximum point of inflection of the curve is annotated. A triangle is formed by drawing the segments from the apex (vertex B) to vertices A and C respectively lying on the NSL. The maximum angle of the spine is calculated as 180 - ABC.

6.3 Study Population

1,929 UK Biobank DXA scans have been annotated, 308 have scoliosis and have been annotated with the maximum modified Ferguson angle using the DXA scoliosis method (DSM) as outlined in [Taylor et al. 2013] and the rest are normals (6 degrees or less). We

use this test set of 308 human angles to evaluate the association between our predicted spine maximum angle and manually annotated angle. For curve location, the Cervico-Thoracic and Thoraco-Lumbar scans have been combined together such that Thoracic contains now Cervico-Thoracic and Thoracic categories (n=138) and Thoraco-Lumbar contains both Thoraco-Lumbar and Lumbar categories (n=175) (see Table 6.1). Among the 1,929 annotated sets, 87.7% have been annotated as having positioning error.

In addition, we have overlap scans from two screening sessions, where 2,728 subjects have two sessions from two screening sessions a year apart. There is no overlap of the two screening sessions with the set of 308 annotated by humans for maximum angle nor the full 1,929 annotated set. The dataset consists of whole-body DXA from the UK Biobank in a standard supine position. All scans are height normalised and resized to consistent resolution (832×320). This resizing operation keeps the aspect ratio, and thus does not affect the measurement accuracy of spinal features in this work.

Table 6.1 presents the descriptive statistics for the curve patterns annotations.

	Annotated Set for Human Angles (n = 308)	
	Individual	**Combined**
Curve Type		
C-shape	286 (92.86)	
S-shape	22 (7.14)	
Curve Location		
Cervico-Thoracic	1 (0.32)	133 (43.18)
Thoracic	132 (42.86)	
Thoraco-Lumbar	41 (13.31)	175 (56.82)
Lumbar	134 (43.51)	
Curve Direction		
Right	171 (55.52)	
Left	137 (44.48)	

Table 6.1: **Curve patterns descriptive statistics.**

6.4 Methods

We introduce the automated DSM method and the validation of our automated spine angle measurement pipeline against the annotated set by humans. From the estimated spine curve, we can measure curve patterns i.e. direction, location, and type.

We show that our model accurately detects scoliosis with high correspondences with

human angle. We propose a simple classifier using features from whole-body angles obtained from geometry to measure positioning error. We further validate our spine maximum angle measurement between the two screening sessions taken a year apart where we assume that curves do not vary much in this time interval in the adult cohort of the UK Biobank.

6.4.1 Spine to Spline: DXA Spine Segmentation and Spline Fitting

We aim to detect scoliosis on DXA scans from the UK Biobank. We use a segmentation network leveraging pseudo-labels from the Avon Longitudinal Study of Parents and Children (ALSPAC) (implementation details available in Appendix A).

As opposed to discrete angle measurements using the standard manual procedures, our DXA curvature method provides angle values on a continuous scale [Jamaludin et al. 2019a]. We perform automatic mask segmentation of the spine and 5 other body parts i.e. head, pelvis, cavity, left leg and right leg in order to obtain a precise delineation of the spine region (see Figure 6.2). Then, we obtain spine midpoints using the spine probability map output from the segmentation model (see Appendix A for more details of the segmentation model). Each midpoint along the spine is computed with the weighted arithmetic mean of the probability and the indices of the scores. We fit a cubic spline through the midpoints extracted from the mask of the spine and we measure the absolute maximum curvature for each single spine curve using derivatives of the spline.

To measure the angle from the spine, we fully automate the DSM method using the spine geometry.

This work introduces the automated DSM from [Taylor et al. 2013], which is a modified-Ferguson method for scoliosis detection tailored to DXA scans as opposed to the widely used Cobb angles on X-rays. The goal is to automate the procedure and obtain the maximum angle of the spine. We use the spine segmentation in Figure 6.2 to obtain a spline. Given the curve of the spine, we show we can automate the human procedure of maximum angle computation. The automated DSM method for angle computation has the advantage of mimicking the human procedure of maximum human angle using the geometry of the spine and outputting directly the angle value without relying on finding a mapping function from curvature to angle. First, the top and bottom Normal Spine Line (NSL) markers are manually determined (see distribution in Figure 6.4). Then, the

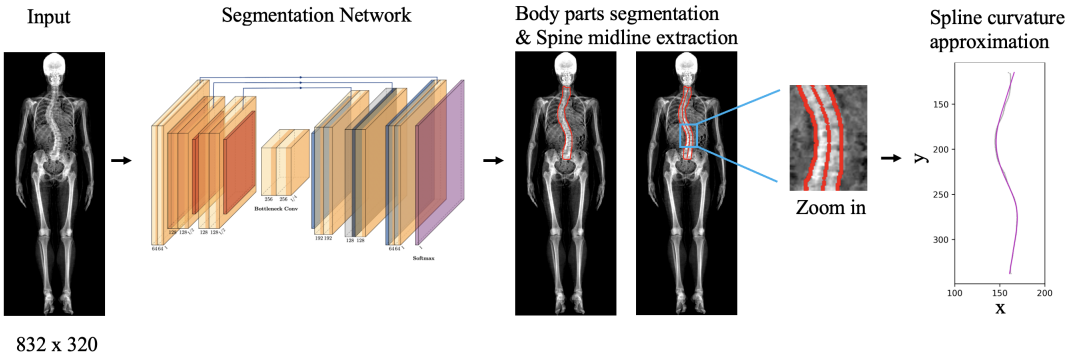


Figure 6.2: **Pipeline of the automated scoliosis measurement method.** We feed into the segmentation network DXA scan (832×320) and ground-truth masks for 6 body parts: head, spine, pelvis, cavity, left leg and right leg. At each row of the spine probability map, the weighted arithmetic mean of the probability and the indices of the scores is calculated to be the predicted midpoint. Cubic spline approximation is finally employed to filter out noisy predictions.

apex of the curve is estimated. By joining those markers, we obtain a triangle as shown in **Figure 4**. From the triangle, the angle at the apex can be determined from standard trigonometry.

6.4.2 Using the Spline for Automated DSM Spline to Angle (Mid Spine Curve to Maximum Modified Fergusson Angle)

This work introduces the automated DSM (Figure 6.1) from [Taylor et al. 2013] which is a modified-Ferguson method for scoliosis detection tailored to DXA scans as opposed to the widely used Cobb angles on X-rays. The goal is to automate the procedure and obtain the maximum angle of the spine. We use the spine segmentation in Figure 6.2 to obtain a spline. Given the curve of the spine, we show we can automate the human procedure of maximum angle computation in Figure 6.2.

The automated DSM method for angle computation has the advantage of mimicking the human procedure of maximum human angle using the geometry of the spine and outputting directly the angle value without relying on finding a mapping function from curvature to angle. First, the top and bottom Normal Spine Line (NSL) markers are manually determined (see distribution in Figure 6.4). Then, the apex of the curve is estimated. By joining those markers, we obtain a triangle as shown in Figure 6.1. From the triangle, the angle at the apex can be determined from standard trigonometry.

This procedure of finding the maximum angle of the spine by automating the DSM can

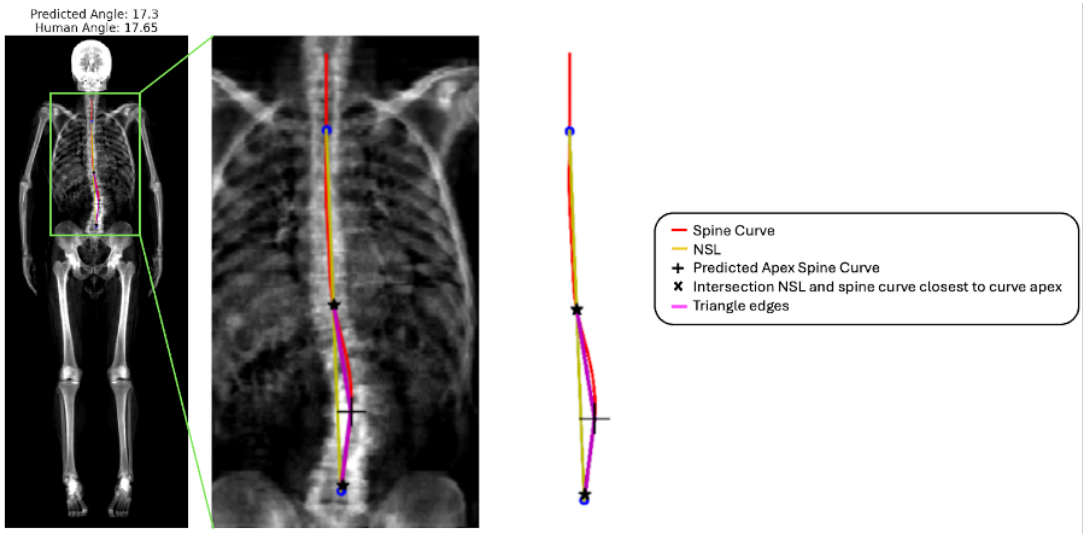


Figure 6.3: **Automated Maximum Angle Computation.** Our automated method reproduces the DSM method to measure the human apex of the spine curve. We draw the lines (magenta) going from the intersection (black stars) of the spine curve (red) and estimated normal spine line (yellow). The angle at the intersection of the magenta segments is the human apex (black cross) of the spine curve.

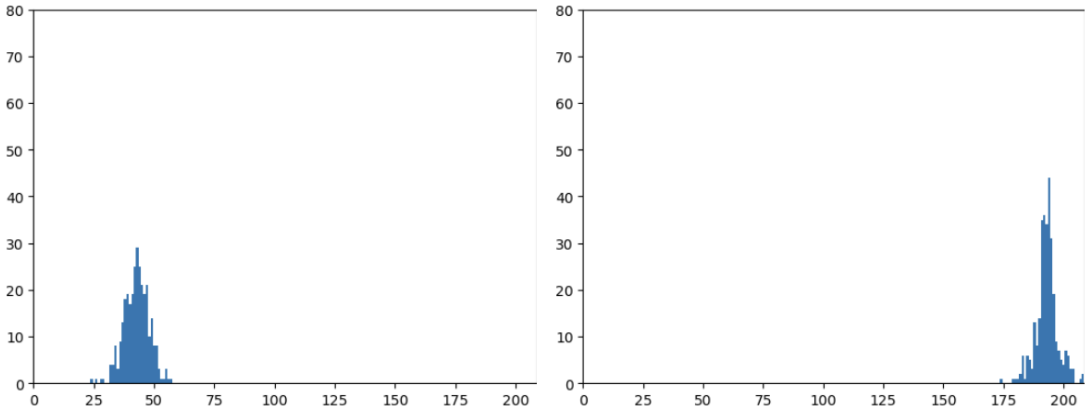


Figure 6.4: **Top (left) and Bottom (right) Normal Spine Line (NSL) Distribution for Human Landmarks.** The dispersion around the mode is greater for the top NSL marker compared to bottom marker which is narrower. This can be explained by the greater difficulty for humans to annotate the point of top of the spine with junction of the clavicle compared to the bottom of the spine given pelvis orientation.

be broken in two steps. First, we estimate the top/bottom coordinates of the NSL. This consists of taking the top and bottom of the spine segmentation by thresholding it to a binary segmentation map and finding the index of maximum and minimum points. Then, we compute the maximum angle of the spine using the triangle shown in Figure 6.3.

6.4.3 Using the Spline for Curve Patterns (Curve Types, Location, Direction)

In this section, we show that given the curve of the spine, we can learn more about the inherent shape of the spine in 2D. This includes right and left curve direction, as well as curve location where the apex is located either in the thoracic or thoraco-lumbar region and the curve types by the number of apex in the curve.

Curve types

The spine curves vary in type depending on the number of extrema. A spine curve with one apex is classified as C-shape curve or equivalently single curve while a spine curve with two apexes is classified as S-shape curve or double curve. Human annotators have annotated the curves with the number of apexes i $\in$ [0:2], i $\in Z$. The challenge is to correctly identify the S-shape curves since the angle values per scan comprise several local extrema. S-shape by definition has two apexes differing in signs. How big in value do the two apexes have to be to avoid local extrema while capturing the relevant apexes ? This is the investigation we show in results curve patterns in Section 6.5.3.

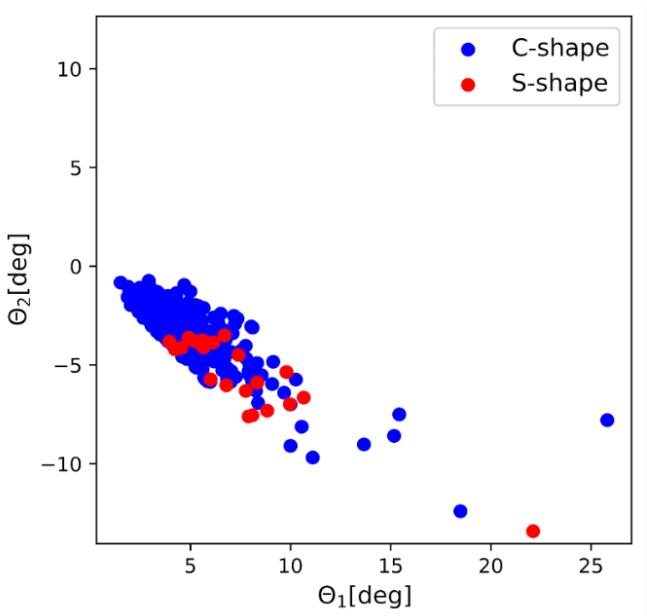


Figure 6.5: **Scatter plot of the maximum angle of the spine θ_1 versus the second maximum apex θ_2.** All θ_1 are made positive, while all θ_2 are made negative for simplicity as we are interested in the magnitude of angle values.

We follow these steps below in order:

- Compute the maximum angle of the spine and take the second largest apex differing

in sign

- Measure the DSM angle at those two apexes
- Measure how big in value the two apexes are
- Define a threshold of minimum value of the two apexes to classify S shape curves from C shape curves

According to the human annotated set of 308 scans, 286 scans have been annotated as having one curve or C-shape, and 22 as double curves or S-shape. Let's denote as θ_1 the maximum angle of the spine and θ_2 the second maximum angle of the spine. For analysis of the correspondences in values, we make sure all θ_1 are positive and θ_2 are negative. From Figure 6.5, the ratio θ_1 to θ_2 is bigger in magnitude on average for S-shape curves (red dots) than for C-shape curves (blue dots). However, there is a region of overlap of the red and blue dots suggesting there is no clear determination of the number of apexes in the curve by considering the maximum and second largest apex in the spine curve only.

Curve location

The location of the curve can be determined using the y index of the spine curve. Clinicians have annotated the location of the spine as either cervico-thoracic, thoracic, thoraco-lumbar and lumbar regions. Given the small number of annotated cervico-thoracic and thoraco-lumbar scans and the challenge to visually identify accurately those locations, we combined those regions in the thoracic and lumbar regions respectively. In our settings, our model aims to predict the thoraco-lumbar region which encompasses the bordering region between thoracic and lumbar and the lumbar region (see Figure 6.6). We measure the location of the spine maximum apex by estimating a dynamic threshold value τ on the vertical axis optimised to separate the thoracic from the thoraco-lumbar region. We opted for binary classification since few of the annotated sets are effectively in the thoraco-lumbar region and this region is challenging to delimitate by humans (see Table 6.1).

Curve direction

In this section, our goal is to automatically determine if the spine curves to the right (right curve direction) or if the spine curves to the left (left curve direction). Right curve

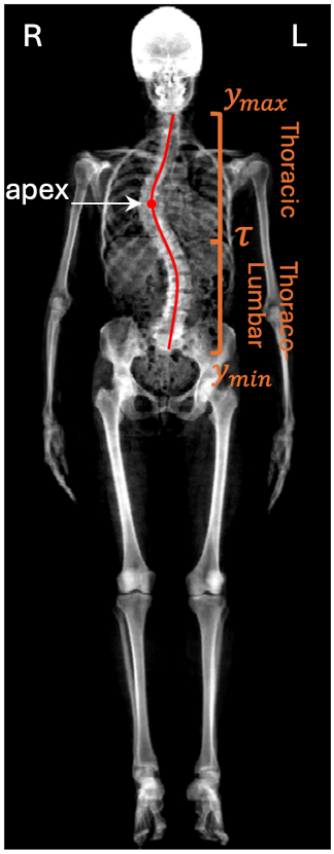


Figure 6.6: **Curve Pattern Explanatory Figure.** We show the maximum apex (white arrow), and the patient right (R) and left (L) side. The direction is determined by the location of the apex pointing to the right or left of the patient. This corresponds to the sign of the curvature of the curve. We present curve location in terms of threshold τ to distinguish the thoracic and thoraco-lumbar regions using spine height values in $[y_{min}, y_{max}]$.

direction corresponds to the left side of the patient while left curve direction corresponds to the right side of the patients in the AP DXA we are showing in Figure 6.6. It is still not yet fully understood why some patterns of curves are generally present in scoliosis cohorts. Adolescents often present right thoracic curves while the apex of lumbar curves tend to be left directed [Castelein 2012]. It was suggested that pre-existent rotation in the normal spine via the process of decompensation could lead to right thoracic and left lumbar curves [Kouwenhoven et al. 2006]. Degenerative scoliosis tends to be left lumbar [Tribus 2003]. This degenerative scoliosis is often displayed by a C-shape curve. We also observe in our cohort of adults a predominance of left lumbar C shape curves (see Table 6.1).

6.4.4 Positioning Error

During scan acquisition, the patient may not be lying still and consequently bending on one side. This lateral bending is hard to control in practice and can affect the measurement of angle and curvature. We show examples of positioning error in Figure 6.7. In the 2K annotation set, 1,693 DXA scans have been labelled as having clear positioning error. In the UKBiobank, a large proportion of scans have positioning error (87.7%) according to the sample of 1,929 DXA scans annotated by humans. The positioning error typically leads to a cervico-thoracic curve. In most cases, the shoulders or head are not aligned and we can detect the scans with positioning error by measuring the angle of these body parts.

According to the definition of positioning error by [Taylor et al. 2013], with the automatic DSM method for maximum human angle measurement, features such as pelvic obliquity or raised shoulder were classified as definite positioning error. Further visual inspection of scans were performed where curves were not clearly explained by poor body positioning. Then, in the next stage, scans were further classified as possible positioning error when it was impossible to classify the curve as either likely scoliosis or definite positioning error. Such cases included curve features that could explain scoliosis but some clearly visible body-positioning errors were identified. To automatically assess body positioning error in the cohort of the UK Biobank, we propose a geometric method measuring angles for head alignment, shoulders alignment, spine endpoints alignment, pelvis alignment and leg symmetry (see Figure 6.7). We trained a classifier with a single linear layer on the set of 1,929 annotated scans for positioning error.

The 5 main input features we give to the classifier are the following:

- **Head Angle**: computed by finding the angle between the line of best fit through the head midpoints and the vertical line
- **Shoulder Angle**: computed by finding the angle between the line through the first non-zero shoulder row and the horizontal line
- **Spine Angle**: computed from the line joining the endpoints of the spine and the vertical line
- **Legs Symmetry**: computed from the lines through knees midpoints after horizontally flipping one knee to measure symmetry

- **Pelvis Angle**: computed by finding the angle between the horizontal line and the top and of the pelvis. We segment the pelvis using pre-trained SAM by setting a bounding box around the pelvis region and fixing negative seeds (5 or less) manually to guide the segmenter [Kirillov et al. 2023]. The top line of the pelvis is obtained by joining two points, top y index of the left and right part of the pelvis mask.

The model is expressed by:

$$Z = W^T X + b$$

where X, corresponds to the input feature vector fed to the neural network, b is the bias term. For each input feature X, there is a corresponding weight W, which signifies how strongly the effect of the input is on the output. Therefore, we discard the patients with positioning error above a certain threshold computed from output of the classifier after visual inspection of the scans following the DSM method (see Figure 6.7). We take the softmax outputs to rank the scans from lower to higher scores. Then, expert clinicians visualised the scans with hierarchical ranking to select the optimal threshold with uncertainty.

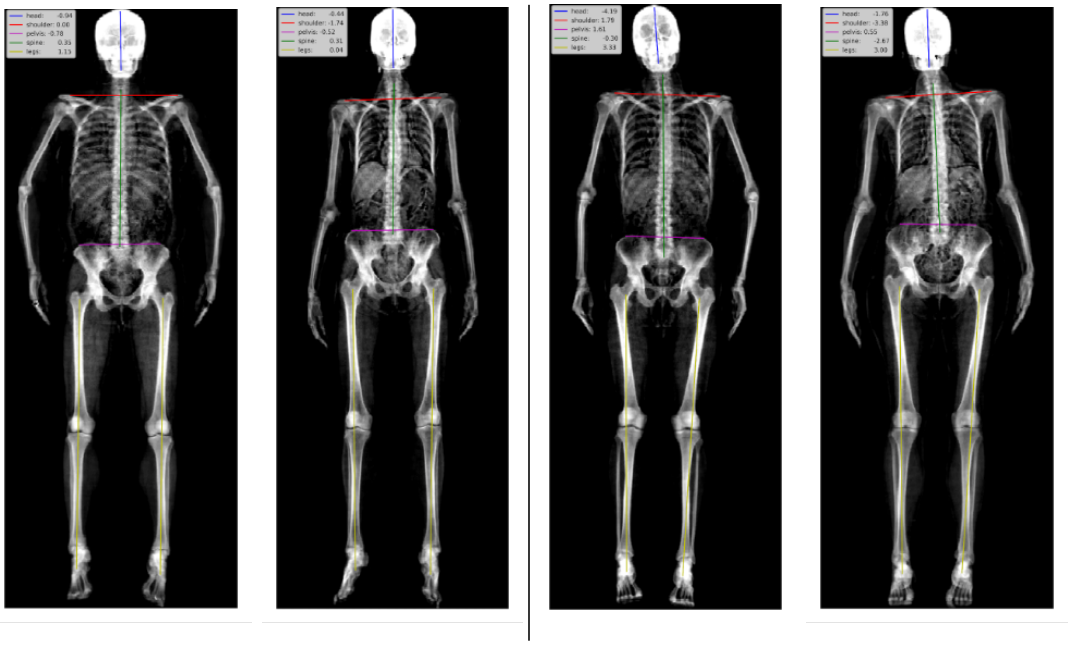


Figure 6.7: **Visualisation of body angles and body positioning error**. Visualisation of body angles and body positioning error. We show two cases of low positioning error < 0.25 (left) and high positioning error > 0.9 (right) with corresponding angles for head, legs, shoulders, spine and pelvis on top of the scan with lines used to draw the angle.

Another explanation to the discrepancy between model output and human annotation are

scans with positioning error. As mentioned earlier, 87.7% of the scans in the UKBiobank have positioning error.

6.5 Results

6.5.1 Automated DSM

The angles obtained from the automated DSM given the segmented spine from Section 6.4.1 show good agreement with human annotation for spine curves varying in severity (see Figure 6.8).

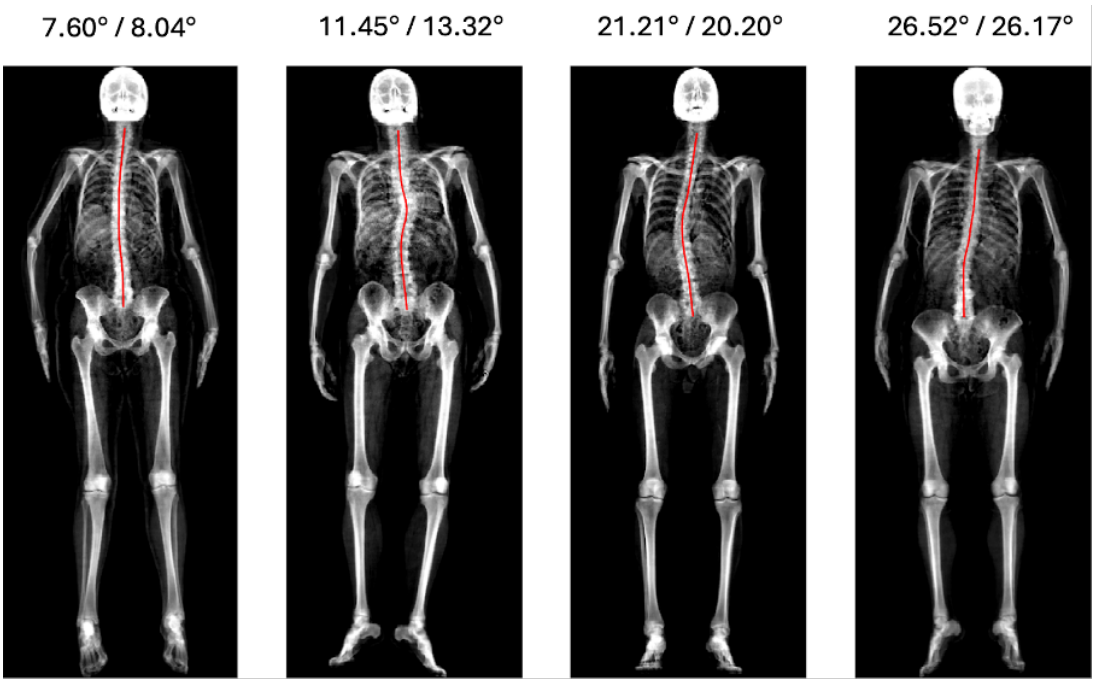


Figure 6.8: **Automated DSM Visualisation of Scoliosis Severity.** We show splines of the spine (red) overlayed on DXA scans for angles ranging from mild (left) to severe (right). On top of each scan we show (i) predicted angle DSM (degree), and (ii) human maximum angle annotated by expert clinicians.

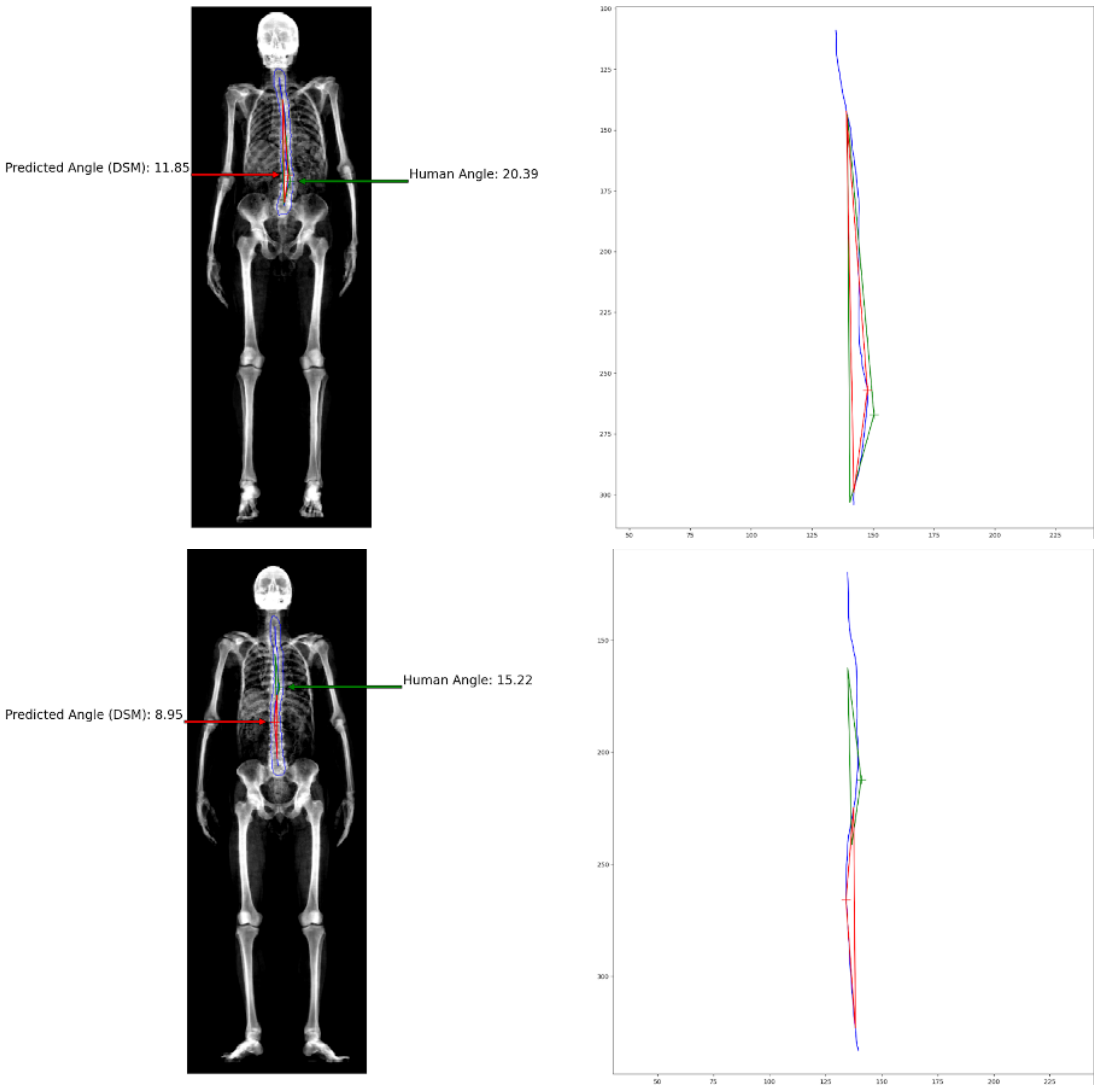


Figure 6.9: **DSM Automated Angles.** We show examples of automated DSM angles from estimated triangles with red for prediction and green for human ground-truth. First, the spine is segmented (blue contour) and the apex of the curve computed using the pipeline in Figure 6.2. The top and bottom vertices of the triangle are estimated relative to the height of the patients, see explanatory diagram in Figure 6.3.

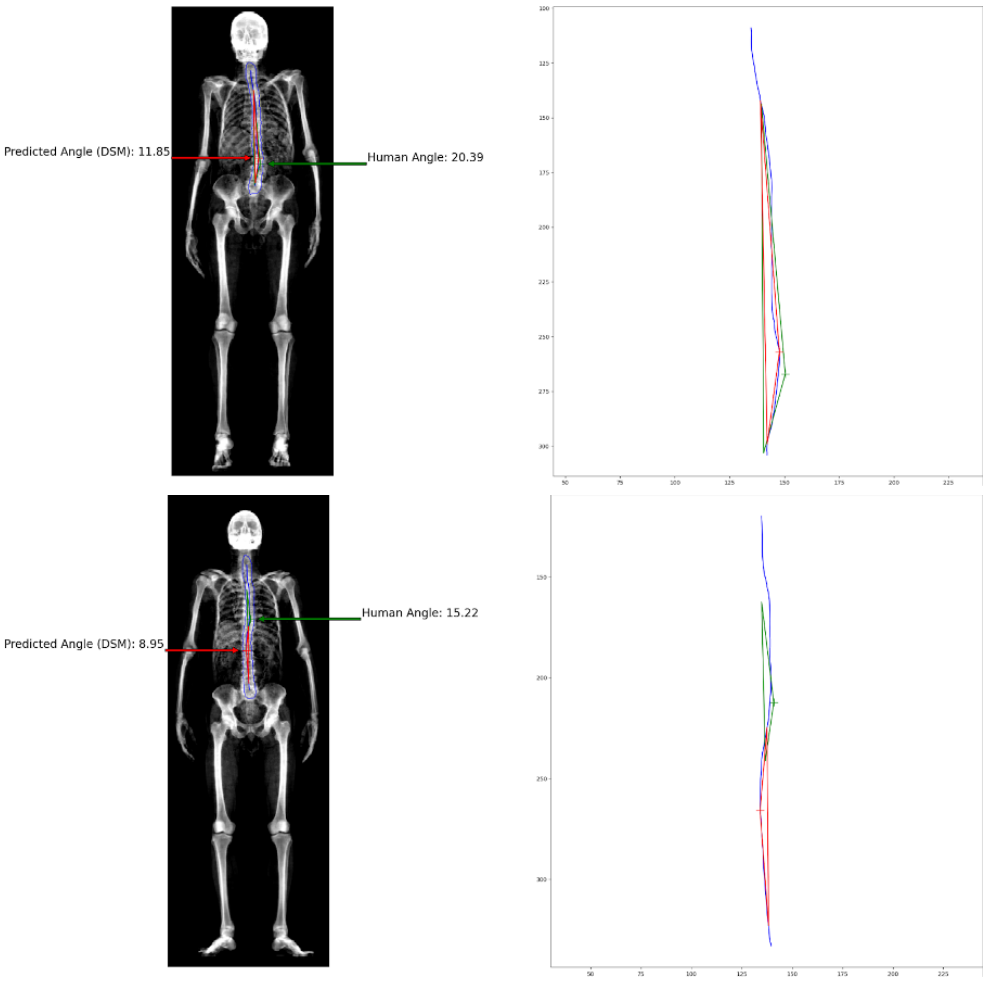


Figure 6.10: **Failure cases DSM Automated Angles.** We show examples of discrepancies between automated DSM angles and human angles. The triangle in red is for prediction and green for human ground-truth. The first example is a failure to estimate the apex location. The second example is a double curve case where our automated DSM did not pick the second largest peak in the curve corresponding to human annotation. The third and fourth examples are wrong estimations of point A and C of the triangle as defined in Figure 6.1. A deviation of vertices for a few pixels leads to a big difference in angle measurements from automated to human method for angles.

We show results of the automated DSM method with comparison with human measurements in Figure 6.9 and 6.10. Failure cases do occur, mainly for small curves where discrepancies between the automated and human measurements are bigger (see Figure 6.10). Our predicted angle values agree with human values with Pearson's R of 0.88 (see Figure 6.11). The mean difference between predicted and human angle values is in the range [-7.73, 1.93] in degrees with 95% confidence (see Figure 6.11). The DSM method (manual) has previously been shown to be reliable where 95% of repeat measures were within 5°. This suggests that our DSM method preserves the error range by humans of 5° error and therefore suggests that our predictions show good agreement with humans. We show some cases of failure of approximation of one of the vertices and

the angles associated in Figure 6.10.

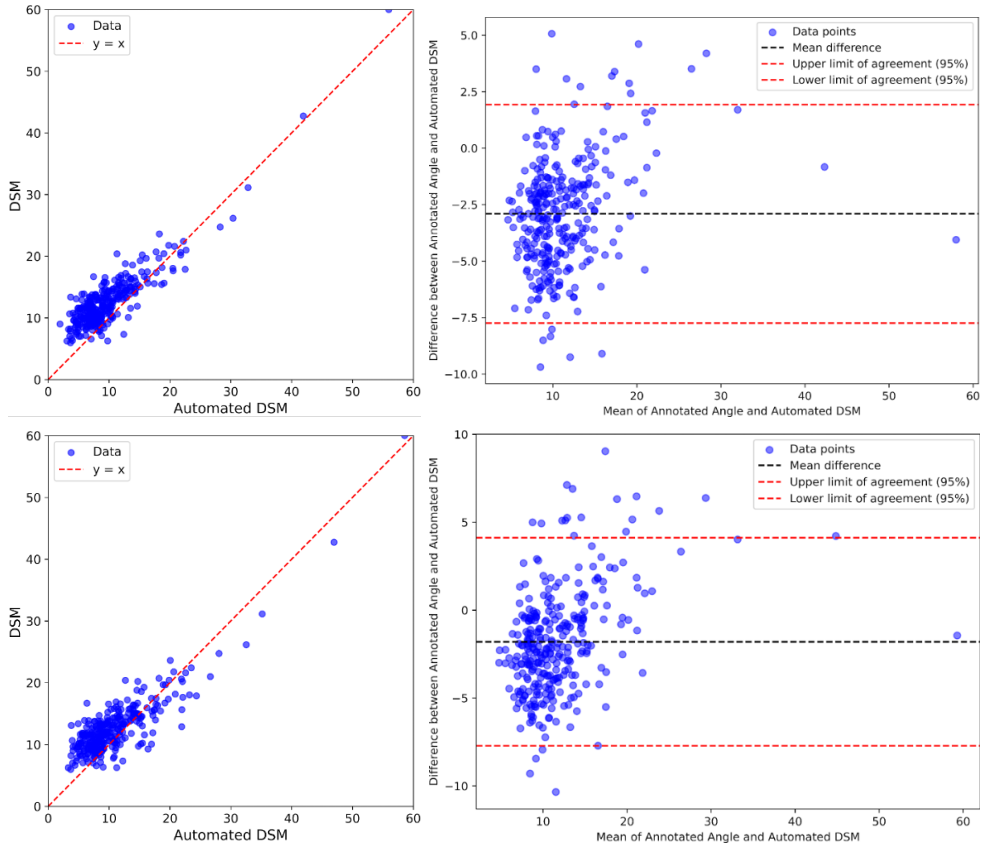


Figure 6.11: **Angle Agreement with Humans**. Left plot compares DSM annotated human angles versus fully automated DSM (ρ = 0.89, n=308). Right plot shows Bland-Altman of Predicted versus Human DSM Angles (degree) with mean difference of -2.90/ median diff:-2.87 and 95% CI of [-7.73, 1.93]. (2nd row) Same as 1st row with additional shifting of apex (0.25px) to the direction of maximum slope of spine curve relative to normal spine line. Left correlation plot has ρ = 0.86 and right Bland-Altman has mean difference of -1.80 /median diff: -2.11 and 95% CI of [-7.72, 4.12].

6.5.2 Verification of normals.

We also compute the DSM angles on the 1,621 set of DXA annotated (1,929 minus the 308 that have scoliosis and were annotated for modified Fergusson angle by [Clark et al. 2014]. This set corresponds to scans with the maximum angle of the spine below the 6 degree threshold defined by humans. We report the distribution of predicted DSM angles with our method scans in the histogram below (see Figure 6.12). The number of scans with predicted DSM angles above 6 degree is n = 431 (26.59%). On average, we have n = 61 (3.76%) above the scoliosis threshold using the underprediction of -2.87 degrees with human annotation.

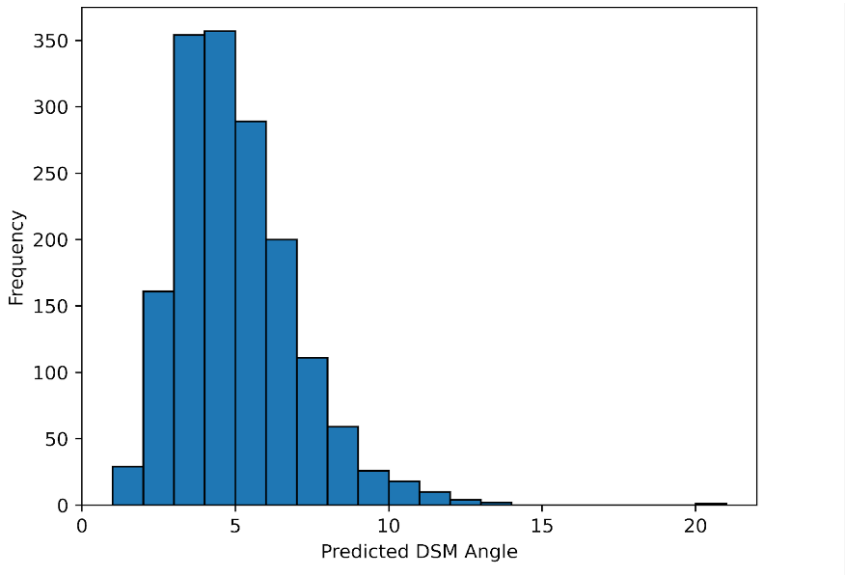


Figure 6.12: **Distribution of DSM for normals (n=1,621)** Histogram showing the distribution of predicted DSM angles with our method on the annotated set of normals (not annotated for DSM angle i.e. modified Fergusson angle).

6.5.3 Curve Patterns (Curve Types, Location, Direction)

Curve Location.

For curve location prediction, our automated model achieves a balanced accuracy of 0.832, specificity of 0.883, sensitivity 0.587, PPV 0.875 and NPV 0.604 (see Figure 6.13).

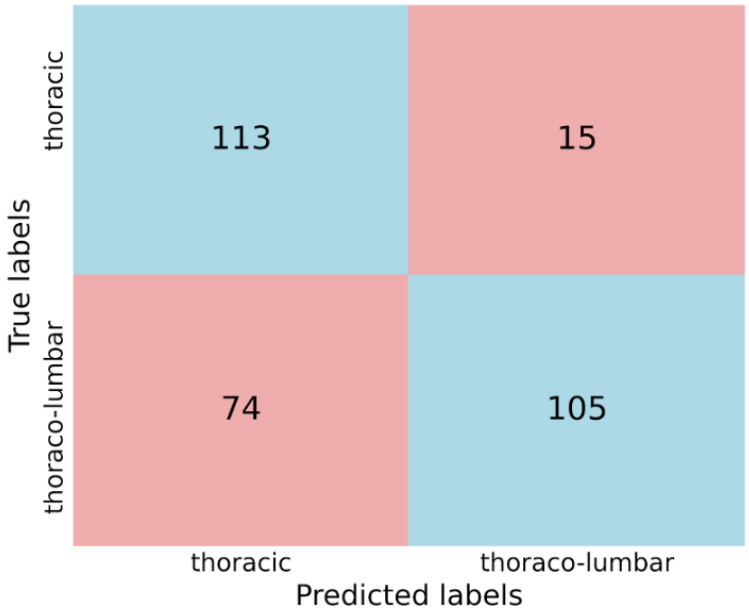


Figure 6.13: **Confusion Matrices Location (Thoraco-lumbar and Lumbar) of curves.** Note that thoraco-lumbar class encompasses the original two categories, thoracic and thoraco-lumbar curves.

Curve Direction.

Figure 6.14 is our performance for curve direction. For the curve direction prediction, our automated model achieves a balanced accuracy of 0.729, Specificity of 0.779, Sensitivity of 0.679, Predictive Positive Value (PPV) of 0.786 and Negative Predictive Value (NPV) of 0.669.

There are ambiguous cases where our method fails. These are mainly cases with close in values. We observe that failures mainly occur for small curves, where the two main apexes (absolute value) are close in value (< 0.01 difference). These cases close in angle values but differ in sign. This also explains our failures on the location task with balanced accuracy of 0.891 (discarding small curves) vs 0.724.

Second potential explanation to the discrepancy between model output and human annotation is human error. Given that scans were manually annotated for right and left by visual inspection, there could be errors. In particular double curve scans or low curve scans are challenging for the human eye to detect. We send the ambiguous cases back to clinicians for manual revision.

Curve Types.

Our objective is to distinguish among three curve types: normal (0 apex), C-shape (1-apex), and S-shape (2-apex). To do this, we employ a simple classifier composed of a single linear layer, following the approach described in section 6.4.4 for the positioning error experiment. Specifically, the classifier distinguishes between C-shape and S-shape curves using the absolute values of the maximum and second-largest apex from the spine curve. To ensure robust training, we stratify the dataset into training, validation, and test sets with an 80-10-10 split. The training set contains 246 entries (with 228 annotated as C-shape and 18 as S-shape), while the validation set includes 31 entries (29 C-shape and 2 S-shape), and the test set comprises 31 entries. We train the classifier for 10,000 epochs using the Adam optimizer at a learning rate of 0.01 and optimize with the Cross Entropy loss function. To address the class imbalance, we weigh the classes in the loss function. The performance of our curve type prediction model is evaluated on the test set, where we achieve an AUC of 61.1%.

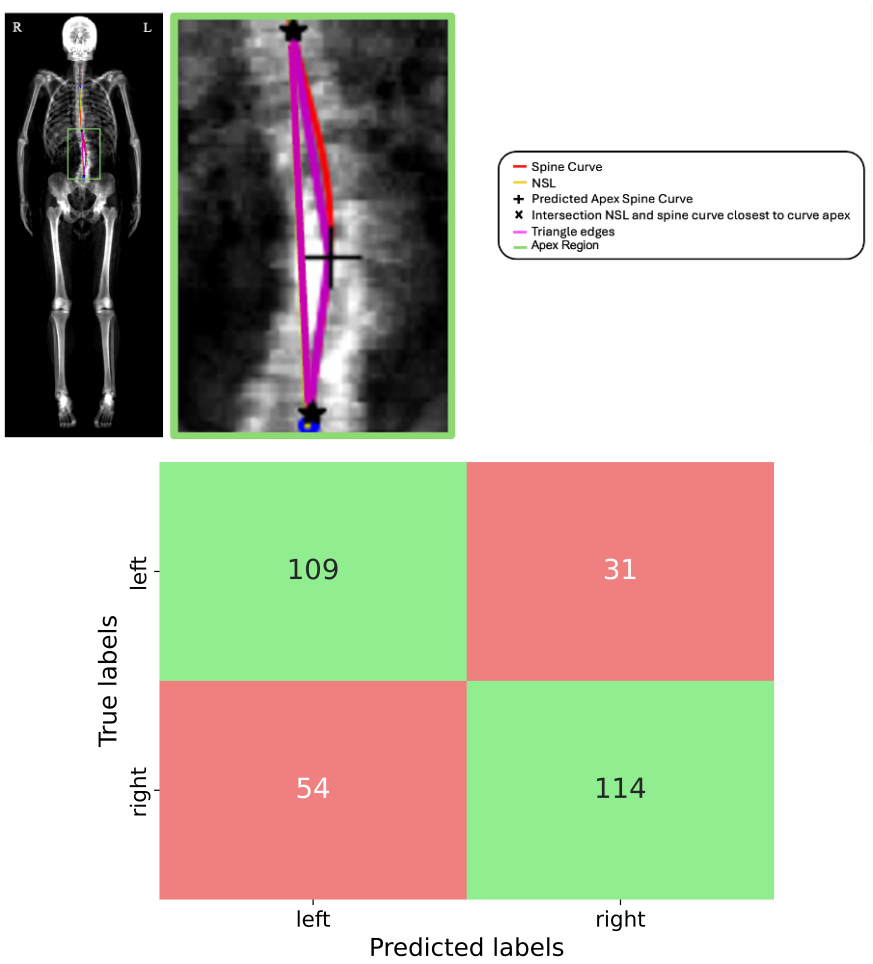


Figure 6.14: **Curve direction prediction.** Top is a diagram showing the apex of the curve (black) and triangle (magenta) within the apex region (green). Bottom is a confusion matrix for curve direction. We show the predictions of the model for right/left against labels on the set of 308 annotated for human angle. We re-run the evaluation with the cases of angles close in value. We check whether the second apex is in the apex region (green region) defined by inflection points on the spine curve (black stars) and closer in distance to the human apex. We obtain a balanced accuracy of 0.92. If we include in the evaluation the second largest apex, the accuracy of our method jumps from 72.4% to 92% suggesting that a significant part of failures from our models are due to small curves with similar magnitudes but differing in signs.

6.5.4 Validation of Spine to Spline and Curve Patterns on Two UKBB Sessions

In the UKBiobank, the same volunteers are scanned at two time points less than a year apart, which provides a unique opportunity to validate our method. Because these sessions are so closely spaced, we expect minimal changes in the spine maximum angle especially in adulthood when growth has ceased and scoliosis progression slows significantly. While it is known that in mature adults scoliosis curves can still progress at a rate of one to three degrees per year, the short interval between sessions should result

in only minor variations in angle values, with the apex of the spine remaining largely consistent. In our evaluation, we observe a strong agreement between the two sessions, with a mean difference of just 0.167° between Session 2 and Session 1 (see Figure 6.15 for the Bland-Altman analysis, which reports a 95% confidence interval around the mean difference of [-3.934, 4.267]). Furthermore, we assess the correspondence in curve patterns as additional validation. For both sessions, the percentage distribution of curve directions is nearly identical: in Session 1, 58.11% of curves were left and 41.89% were right, compared to 58.07% left and 41.93% right in Session 2. Similarly, the location of the apex is preserved across sessions. In Session 1, 39.57% of the apexes were thoracic and 60.43% thoraco-lumbar, which is comparable to the 39.49% thoracic and 60.51% thoraco-lumbar distribution observed in Session 2. Although the size of the curves might evolve slightly over this short period, the number of apexes is very stable. Both sessions report 92.31% C-shaped curves and 7.69% S-shaped curves.

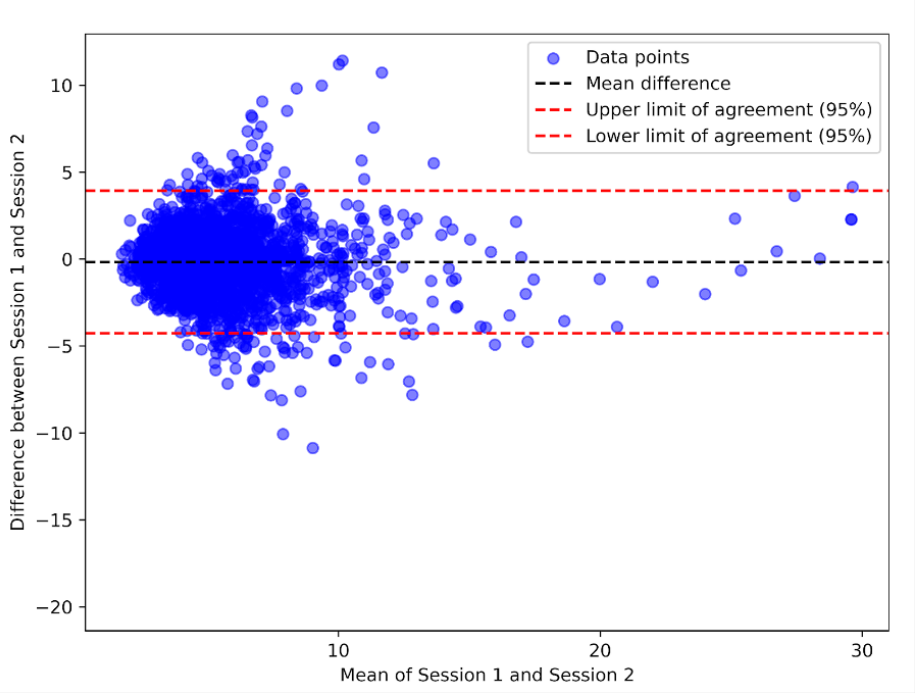


Figure 6.15: **Bland-Altman between two Sessions.** Mean difference is 0.167 and 95% confidence interval around the mean difference is [-3.934,4.267].

6.6 Conclusion

This study introduces and validates an automated method for measuring scoliosis parameters from Dual-energy X-ray absorptiometry (DXA) scans in a large adult cohort from the UK Biobank. Our method, based on a geometric approach using spline fitting,

demonstrates strong agreement with human annotations, achieving a Pearson's correlation of 0.87 and a mean difference of less than 5 degrees between automated and manual angle measurements. The method also successfully predicts curve location, direction, and type, showcasing its utility in providing a comprehensive assessment of scoliosis parameters. The automated DSM method overcomes several limitations of traditional scoliosis measurement approaches, including the reliance on X-rays, the time-intensive manual process, and the challenge of assessing complex 3D spinal deformities using 2D imaging. By leveraging machine learning and geometric techniques, this approach allows for consistent, reliable, and efficient scoliosis measurement, even in lower-resolution DXA scans. Our findings highlight the potential of automated scoliosis assessment methods in large-scale population studies. The validated pipeline facilitates the identification of scoliosis phenotypes and their associations with other biomarkers, such as age, body composition, and anthropometric measurements. This capability could enhance screening programs, enabling cost-effective, low-radiation, and high-throughput scoliosis assessment, particularly in aging populations where degenerative scoliosis is prevalent. In conclusion, this work represents a significant step forward in the automated assessment of scoliosis in adults, providing a scalable and reproducible tool for research and clinical applications. By enabling accurate and efficient scoliosis measurement, this method has the potential to advance our understanding of scoliosis progression and improve patient care. Future work should integrate multimodal data (genetic, biomechanical, and imaging) to refine predictive models.

A Model Implementation Details

ALSPAC scans are lower in resolution (5 $\times$ lower) compared to the UK Biobank, and ALSPAC participants are all adolescents while the UK Biobank is entirely made up of adult volunteers. We use a pseudo-labelling approach and iterative training to improve the segmentation of the spine and to handle the domain gap between ALSPAC and UK Biobank cohorts. We find that we can achieve better results in terms of reliability by fine tuning on the UK Biobank dataset (see details in [Bourigault et al. 2022]).

Our segmentation model follow the standard U-Net from [Ronneberger et al. 2015]. The input is height normalised DXA scan (832,320) and targets are body masks for head, spine, cavity, right leg and left leg. The output of the segmentation model is (832,320,5) where each of the 5 channels corresponds to mask output for the body parts. See Figure A2 for details of the resolutions and channels. Qualitative visualisations of the body part segmentation are available in Figure A1.

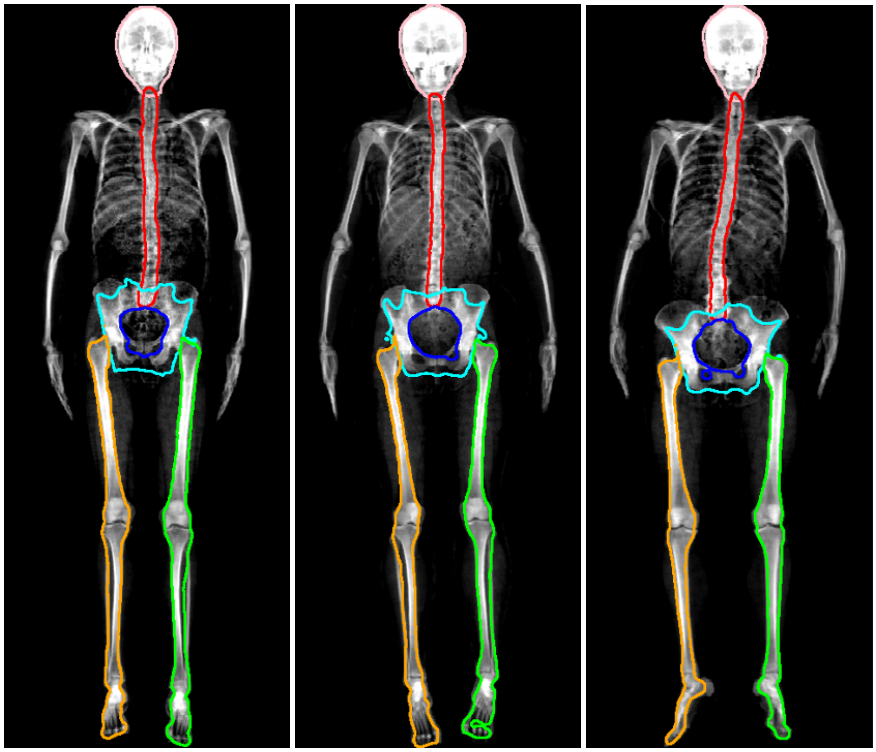


Figure A1: **Body part segmentation for head, spine, cavity, pelvis, left leg and right leg.These segmentation contours were obtained using the U-Net model in Figure A2 from Bourigault E. et al, 2022.**

Layer	Resolution	Channels
Input	832×320	1 (Grayscale)
Convolution Block 1	832×320	64
Bilinear Interpolation 1	416×160	64
Convolution Block 2	416×160	128
Bilinear Interpolation 2	208×80	128
Convolution Block 3	208×80	256
Bilinear Interpolation 3	104×40	256
Convolution Block 4	104×40	512
Bilinear Interpolation 4	52×20	512
Bottleneck	52×20	1024
Up Convolution 1	104×40	512
Convolution Block 5	104×40	512
Up Convolution 2	208×80	256
Convolution Block 6	208×80	256
Up Convolution 3	416×160	128
Convolution Block 7	416×160	128
Up Convolution 4	832×320	64
Convolution Block 8	832×320	64
Output	832×320	5 (Output Channels)

Figure A2: **Our U-Net architecture with layer-wise resolution and number of channels.**

Chapter 7

UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation

This paper is to be published in the proceedings of the International Conference on Computer Vision, 2025 [Bourigault et al. 2025].

This works improved the 3D spine segmentation labels from Chapter 4. A current drawback of medical imaging segmentation models is the small scale of the dataset used for training thus impeding good generalisation to external datasets. Furthermore, the difference in imaging protocols, and modality means that a segmentation model trained on one modality typically fail to generalise for out-of-domain data at inference time. We tackled this issue in this work by leveraging a large dataset of more than 51K MRI from the UK Biobank [Sudlow et al. 2015] and segmentation labels from TotalVibeSegmentator [Graf et al. 2024] that we filtered with our custom body filtration method. We trained a state-of-the-art segmentator model on this filtered dataset and used entropy test-time adaptation to apply it to out-of-domain dataset i.e. abdomen CT and MRI, and brain MRI. We showed improved segmentation of the spine compared to our baseline and manually annotated set, and a good performance on the different public datasets of various organs.

UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation

Emmanuelle Bourigault
emmanuelle@robots.ox.ac.uk

Amir Jamaludin
amirj@robots.ox.ac.uk

Abdullah Hamdi
abdullah.hamdi@eng.ox.ac.uk

Visual Geometry Group, University of Oxford

Abstract

In medical imaging, the primary challenge is collecting large-scale labeled data due to privacy concerns, logistics, and high labeling costs. In this work, we present the UK Biobank Organs and Bones (UKBOB), the largest labeled dataset of body organs of 51,761 MRI 3D samples (17.9 M 2D images) and a total of more than 1.37 billion 2D segmentation masks of 72 organs based on the UK Biobank MRI dataset. We utilize automatic labeling, introduce an automated label cleaning pipeline with organ-specific filters, and manually annotate a subset of 300 MRIs with 11 abdominal classes to validate the quality (UKBOB-manual). This approach allows for scaling up the dataset collection while maintaining confidence in the labels. We further confirm the validity of the labels by the zero-shot generalization of trained models on the filtered UKBOB to other small labeled datasets from a similar domain (e.g. abdominal MRI). To further elevate the effect of the noisy labels, we propose a novel Entropy Test-time Adaptation (ETTA) to refine the segmentation output. We use UKBOB to train a foundation model (*Swin-BOB*) for 3D medical image segmentation based on Swin-UNetr, achieving state-of-the-art results in several benchmarks in 3D medical imaging, including BRATS brain MRI tumour challenge (+0.4% improvement), and BTCV abdominal CT scan benchmark (+1.3% improvement). The pre-trained models and the code are available at https://emmanuelleb985.github.io/ukbob, while filtered labels will be made available with the UK Biobank.

7.1 Introduction

The advent of large-scale labeled datasets such as ImageNet [Russakovsky et al. 2014] and LAION [Schuhmann et al. 2021] has been a cornerstone in the remarkable progress of computer vision, enabling the development of powerful foundation models [Radford et al. 2021; Rombach et al. 2022] that excel across various tasks. These models benefit

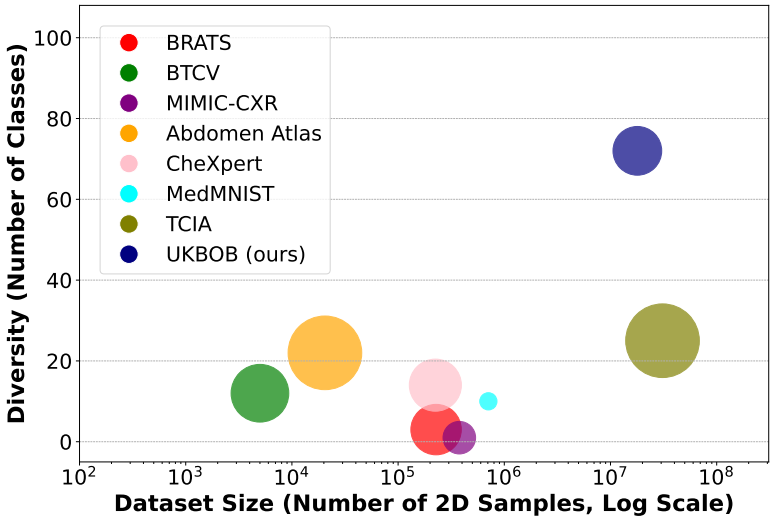


Figure 7.1: **UKBOB Size and Diversity Plot.** Our proposed **UKBOB** is the largest labeled medical imaging dataset for segmentation, largely surpassing the size and diversity of previous datasets in 2D segmentation and 3D segmentation. This new scale in size and diversity should unlock a new wave of applications and methods in the computer vision and medical imaging communities. The size of the bubbles indicates 2D image resolution.

immensely from the abundance of labeled data, which allows them to learn rich and generalizable representations. In stark contrast, the medical imaging domain grapples with a significant scarcity of large-scale labeled datasets due to stringent privacy regulations, complex logistics, and the high costs associated with expert annotations [Willemink et al. 2020; Tajbakhsh et al. 2020; Litjens et al. 2017]. This limitation hampers the development of robust and generalizable models for critical tasks such as 3D medical image segmentation, which is essential for disease diagnosis, treatment planning, and patient monitoring.

Previous medical image datasets either lack diversity or are too small for generalization [Fang and Yan 2020; Baid et al. 2021; Irvin et al. 2019]. Recognizing the pressing need for extensive and diverse medical imaging datasets, we introduce the **UK Biobank Organs and Bones (UKBOB)**, the largest labeled segmentation medical imaging dataset to date. Based on the UK Biobank MRI dataset [Sudlow et al. 2015], UKBOB comprises 51,761 3D MRI scans and over *one billion* 2D segmentation masks covering 72 organs. This dataset not only surpasses existing medical imaging datasets in scale but also in anatomical diversity, providing an unprecedented resource for training robust and generalizable models (see Figure 7.1). Table 7.1 highlights the differences in scope and other aspects between the different datasets. To collect the labels of UKBOB,

we leverage automatic labeling based on the newly released TotalVibe Segmentator [Graf et al. 2024]. However, the automatic labeling of such a vast dataset introduces challenges related to label noise and quality assurance. To tackle this, we propose a novel mechanism for filtering organ labels based on a statistical Specialized Organ Labels Filter (SOLF). We also collect manual labels from 300 MRIs for 11 abdominal organs acting as validation (UKBOB-manual). We further account for noisy labels and dynamically refine the segmentation based on the model's confidence using a novel Entropy Test-Time Adaptation (ETTA). These approaches ensure high-quality labels and enhances the model's robustness. We validate the validity of the labels by demonstrating zero-shot generalization of the trained models on the filtered UKBOB dataset to other datasets from similar domains, such as the AMOS abdomen MRI dataset [Ji et al. 2022] and the BTCV abdomen CT dataset [Fang and Yan 2020].

Leveraging the extensive UKBOB dataset, we train *Swin-BOB*, a foundation model for 3D medical image segmentation based on Swin-UNetr [Hatamizadeh et al. 2022]. Our model achieves state-of-the-art performance on several benchmarks in 3D medical imaging, including the BRATS brain tumor MRI challenge [Baid et al. 2021] and the BTCV abdominal CT scan benchmark [Fang and Yan 2020]. Our contributions can be summarized as follows:

Contributions: **(i)** We introduce UK Biobank Organs and Bones (UKBOB), the largest labeled dataset of organs, consisting of 51,761 MRI 3D samples and a total of 1.37 billion 2D segmentation masks of 72 organs based on the UK Biobank MRI dataset. **(ii)** We leverage automatic mechanisms for cleaning and filtering the labels based on body statistics and specialized organ filter, allowing for high-quality scale-up of the labels. The collected labels are validated by a subset of 300 manually annotated labels of 11 abdominal organs. **(iii)** To further elevate the effect of the noisy labels, we propose a novel Entropy Test-time Adaptation (ETTA) to refine the segmentation outputs. **(iv)** We train *Swin-BOB*, a foundation model for 3D medical image segmentation based on Swin-UNetr network [Hatamizadeh et al. 2022], achieving state-of-the-art results on standard benchmarks in 3D medical imaging.

Attribute	Medical Imaging Segmentation Datasets					
	BRATS [Baid et al. 2021]	**BTCV**[Fang and Yan 2020]	**MIMIC-CXR [Johnson2019MIMICCXRAD]**	**Abd.Atlas** [Li et al. 2024b]	**Total Segmentator** [Wasserthal et al. 2023]	**UKBOB (ours)**
Number of Classes	3	12	1	25	104	72
Number of 3D Samples	1,470	50	N/A	20,460	1,204	**51,761**
Total Number of 2D Images	227,850	5,000	377,110	673,000	400,000	**17,902,080**
Number of 2D Label Masks	581,715	425,000	N/A	16,825,000	5,800,000	**1,378,913,040**
Number of Patients	1,470	50	227,835	N/A	1,204	50,000
Meta Information	N/A	N/A	Text Reports	N/A	N/A	Bone Density + Fat %
Scope	Brain	Abdomen	Chest	Abdomen	Full-Body	Full-Body
Modality	MRI	CT	X-rays	CT	CT	MRI
Specialty	Tumour	Organs	COVID	Organs	Organs/Bones	Organs/Bones
2D Image Resolution (axial)	240×240	314×214	2500×3056	280×280	512×512	224×174

Table 7.1: **Comparison of Different Medical Imaging Segmentation Datasets**. We compare our proposed UKBOB to other well-known medical image segmentation datasets in terms of scope, size, and modality.

7.2 Related Work

3D Segmentation in Medical Imaging. Advancements in deep learning have significantly influenced 3D data processing, leading to various approaches such as point-based methods [Qi et al. 2017a; Qi et al. 2017b], voxel-based methods [Maturana and Scherer 2015-09; Choy et al. 2019], and view-based methods [Su et al. 2015; Hamdi et al. 2021; Hamdi et al. 2023b; Mai et al. 2024; Mai et al. 2023; Hamdi et al. 2023a; Held et al. 2023]. In medical imaging, the U-Net architecture [Ronneberger et al. 2015] revolutionized image segmentation with its symmetric encoder-decoder structure and skip connections, becoming widely used for tasks such as organ segmentation and tumor detection [Qayyum et al. 2017; Pfeffer and Ling 2022; Kirillov et al. 2023]. For 3D volumetric data like MRI or CT scans, 3D U-Net variants have extended this architecture by replacing 2D operations with 3D counterparts, enhancing performance in volumetric segmentation tasks. Recent developments in label-free segmentation utilize self-supervised learning and multimodal foundation models [Zhang et al. 2022; Huang et al. 2023b; Ha and Song 2022; Peng et al. 2023; Kerr et al. 2023; Kobayashi et al. 2022; Ding et al. 2023; Lu et al. 2023; Zeng et al. 2023; Takmaz et al. 2023; Chen et al. 2023; Mai et al. 2023] to segment 3D scenes without explicit labels. However, all of these models require large-scale medical datasets to generlize well [Johnson et al. 2019; Irvin et al. 2019; Rajpurkar et al. 2018; Yang et al. 2023]. While datasets like CheXpert [Irvin et al. 2019] focus on chest imaging with extensive collections of X-ray images and associated clinical labels, they lack detailed segmentation masks necessary for advanced anatomical analysis. Datasets such as AbdomenAtlas-8K and Abdomen Atlas 1.1 [Qu et al. 2023; Li et al. 2024b] provide valuable multi-organ CT scans with organ-level annotations but are limited to specific regions or modalities. The UK Biobank Imaging Study [Sudlow et al. 2015], one of the largest clinical trials, has collected extensive MRI data; however, prior works have not fully leveraged its potential for comprehensive organ segmentation. Our proposed *UK Biobank Organs and Bones (UKBOB)* dataset leverages this resource,

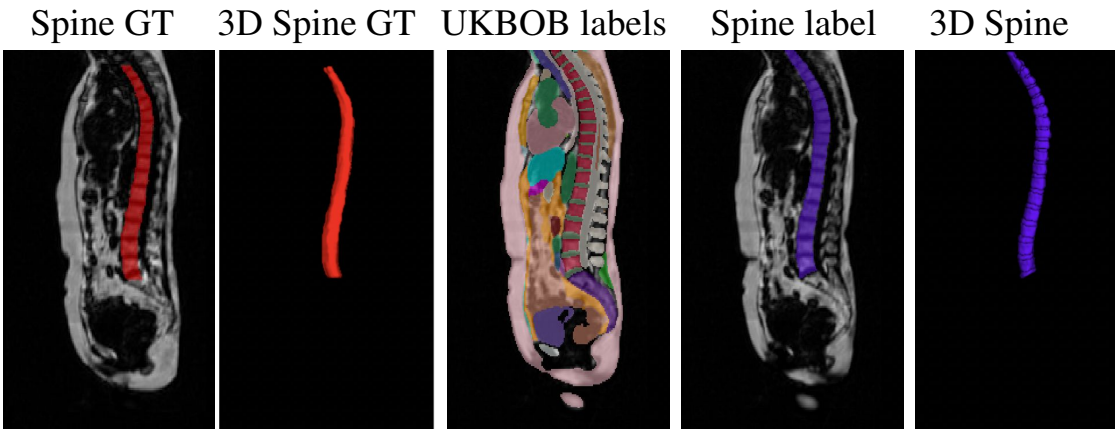


Figure 7.2: **Accuracy of UKBOB Labels.** An example of segmentation labels in UKBOB is shown in the sagittal view. The labels include "*spine*" (in purple) which we can compare to previously collected hand labels of the spine [Bourigault et al. 2023] (in red). We note that the newly collected labels match the manual labels in the spine with a total Dice score of 81.1% on a set of 250 manually annotated test samples, indicating accurate labels.

presenting the largest labeled collection of MRI scans with detailed segmentation masks for 72 organs. By introducing a novel filtering mechanism based on normalized body statistics, we ensure high-quality labels while scaling up dataset collection, enabling the training of foundational models for 3D medical image segmentation with significant improvements over existing benchmarks.

Full-Body MRI Analysis. Most automatic MRI methods have focused on segmenting individual organs or tumors [Chen et al. 2020; Doran et al. 2017; Windsor and Jamaludin 2020; Ranjbarzadeh et al. 2021], with limited research on whole-body scans. Studies that consider full-body imaging often emphasize the spine [Jamaludin et al. 2017; Jamaludin et al. 2018; Windsor et al. 2020; Windsor et al. 2021; Bourigault et al. 2022; Bourigault et al. 2024a], which is crucial for applications like scoliosis detection. Recently, [Graf et al. 2024; Akinci D'Antonoli et al. 2025] released a full torso TotalVibeSegmentator first trained on a subset of NAKO (85 subjects) and UK Biobank (16 subjects) with a nnUNet[Isensee et al. 2021] network. Their network, while useful, does not provide rich enough information to show improved performance on medical image segmentation tasks. Our work builds upon these efforts by using the TotalVibeSegmentator network to collect labels for the 51,761 samples of UK Biobank , filtering MRI labels and verifying their segmentation quality. This enables the training of a general MRI foundation model (Swin-BOB) that can generalize to various tasks and modalities.

Domain Adaptation in Medical Imaging. While U-Net-like networks and their vari-

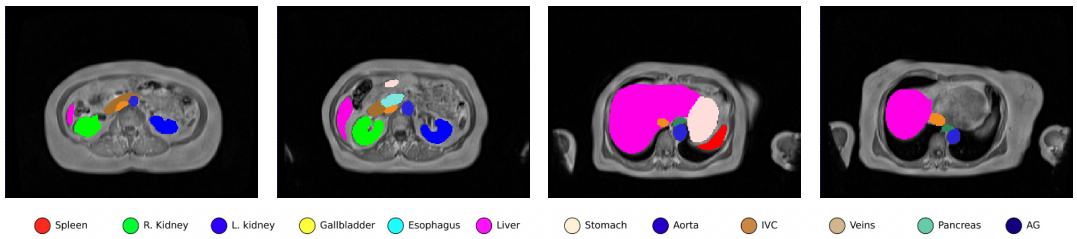


Figure 7.3: **UKBOB-Manual.** We collect manual labels for 300 samples of UKBOB for 11 abdominal organs totaling 3,000 images. UKBOB-manual acts as manual validation for the large UKBOB. Examples of axial slices are shown here.

ants perform well in supervised medical image segmentation, significant performance degradation occurs when the test data differs from training data due to variations in protocols, scanners, or modalities [Ma et al. 2024]. Test-time adaptation (TTA) addresses this by fine-tuning model parameters at test time using only test data without ground-truth [Karani et al. 2021]. Methods like TENT [Wang et al. 2021a] minimize prediction entropy at test time to improve robustness and segmentation performance. Augmentation-based Test-Time Adaptation is proposed [Zhang et al. 2020] to improve on the domain gap issue. However, these approaches rely on well-calibrated models and may be sensitive to augmentation procedures. Recent work [Dong et al. 2024] integrates different predictions using various target domain statistics to enhance performance. Our work tries to address the issue of domain gap when the training domain has noisy labels making the adaptation even more challenging. Our model utilizes the confidence in predictions to adapt based on the entropy map on the test samples. This increases the model's robustness across a wider domain gap.

7.3 Methodology

7.3.1 UKBOB Dataset Labels Collection

UKBiobank is a comprehensive dataset of 51,761 full-body MRIs from more than 50,000 volunteers[Sudlow et al. 2015], capturing diverse physiological attributes across a broad demographic spectrum. This dataset is unlabeled, which limits the potential applications for medical image understanding. We construct the UKBOB dataset by leveraging the UK Biobank MRI Study [Sudlow et al. 2015], which consists of 51,761? neck-to-knee 3D MRI scans. Each scan includes four sequences: fat-only, water-only, in-phase, and out-of-phase images. To obtain segmentation labels for $C = 72$ organs, we employ the TotalVibeSegmentator [Graf et al. 2024], an automatic segmentation tool trained on a subset of UK Biobank data. This approach allows us to generate over 1.37 billion 2D

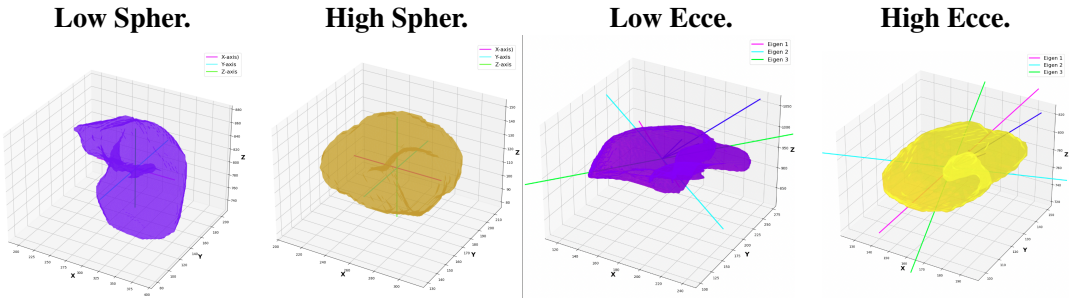


Figure 7.4: **Specialized Organ Label Filter (SOLF).** SOLF integrates sphericity, eccentricity, and normalized volume to statistically filter out inaccurate organ labels. From left to right, the panels display examples of low sphericity (0.21), high sphericity (0.95), low eccentricity (0.14), and high eccentricity (0.87).

segmentation masks. Automatic labeling at this scale is crucial due to the impracticality of manual annotation. While it is not feasible to confirm the quality of 17.9M annotated images manually, we describe next a robust quality control mechanism on the collected labels to insure accurate labels.

7.3.2 Organ Labels Quality Control

Manual Labels for Verification. We design a mechanism to validate these collected labels by humans. To do so we collect manual labels from 3000 2D image from 300 MRI samples for 10 abdominal organs (UKBOB-manual). These manual labels (see examples in Figure 7.3) act as a validation for the large UKBOB dataset. On these labels, the UKBOB automatic labels obtain an average Dice Score of 0.891 (see Table 7.3). Furthermore, we verify the spine labels of UKBOB using previously collected manual labels of 200 3D spine labels [Bourigault et al. 2023]. We show an example in Figure 7.2 and we see how the new collected labels match the manual labels in the spine with a total Dice score of 0.811, indicating accurate labels. We discuss in Section 7.5.1 another mechanism for verifying the labels by zero-shot generalization of trained models to other similar datasets that has manual labels.

Specialized Organ Label Filter (SOLF). While automatic labeling enables creating large datasets, it introduces the possibility of noisy or erroneous labels. To mitigate this, we propose a filtration mechanism that removes outliers from segmentation. A question arises on how to distinguish segmentation failures from common patient abnormalities, e.g. enlarged liver. It's important to note that human organs follow typical geometric properties that arise from the body's need to optimize function while minimizing energy expenditure and structural stress. They reflect the underlying biological "blueprint" that

has been honed by evolution [Shetty et al. 2023]. inspired by the evolutionary regularity of human organs [Shetty et al. 2023], we propose the Specialized Organ Label Filter (SOLF), using three features *jointly*: *normalized volume*, *eccentricity*, and *sphericity* (illustrated in Figure 7.4).

For each organ class c, the normalized volume for some 3D sample is computed as $v_c = \frac{V_c}{V_{\text{body}}}$, where V_c is the voxel count for the organ $c \in \{1, 2, \ldots, C\}$ and V_{body} is the total body voxel count. We define acceptable bounds for each feature by excluding the extreme ϵ percentiles. For example, the bounds for volume are set as

$$v_c^{\text{min}} = P_{\epsilon/2}\left(\{v_c\}_{n=1}^{N}\right), \quad v_c^{\text{max}} = P_{100-\epsilon/2}\left(\{v_c\}_{n=1}^{N}\right) \tag{7.1}$$

, where $P_p(\cdot)$ denotes the p-th percentile function. Sphericity is defined as $\Phi_c = \pi^{1/3}(6V_c)^{2/3}/A_c$, with V_c computed from voxel counts and A_c as the surface area measured by counting the exposed voxel faces of the organ. Finally, eccentricity is defined as $E_c = \sqrt{1 - \lambda_{\text{min}}/\lambda_{\text{max}}}$, where λ_{min} and λ_{max} are the smallest and largest eigenvalues of the covariance matrix of organ c voxel coordinates. A sample is flagged as `inaccurate` if at least two of the three features (normalized volume, eccentricity, and sphericity) fall outside their respective acceptable ranges. Setting ϵ for SOLF effectively discards samples with anomalous organ characteristics while retaining valid labels. A single patient with abnormal organs is extremely unlikely to have more than a single independent aspect of deviation *at the same time*, hence indicating inaccurate labels.

We filter collected organ labels using the *patient's full-body statistics*, a novel approach compared to previous methods that rely on flat label statistics (IQR) [Cheng et al. 2025; Kuş and Aydin 2024] rather than patient meta-information and organ-specific features .

7.3.3 Entropy Test-Time Adaptation (ETTA)

A common practice in medical imaging to addresses the labels distribution-shift is to employ Test-Time Adaptation (TTA) by fine-tuning model parameters at test time using only test data without ground-truth [Karani et al. 2021]. To mitigate the impact of any residual label noise at UKBOB, we introduce Entropy Test-Time Adaptation (ETTA). ETTA refines the model's predictions on test samples by fine-tuning the batch normalization parameters using the test data itself, guided by minimizing the prediction entropy. Given a test sample $\mathbf{x}$, we first obtain the network's initial prediction $\mathbf{p} = f_\theta(\mathbf{x})$, where $\mathbf{p} \in [0, 1]^{N \times C}$ is the softmax probability over C classes at each of the N voxels in

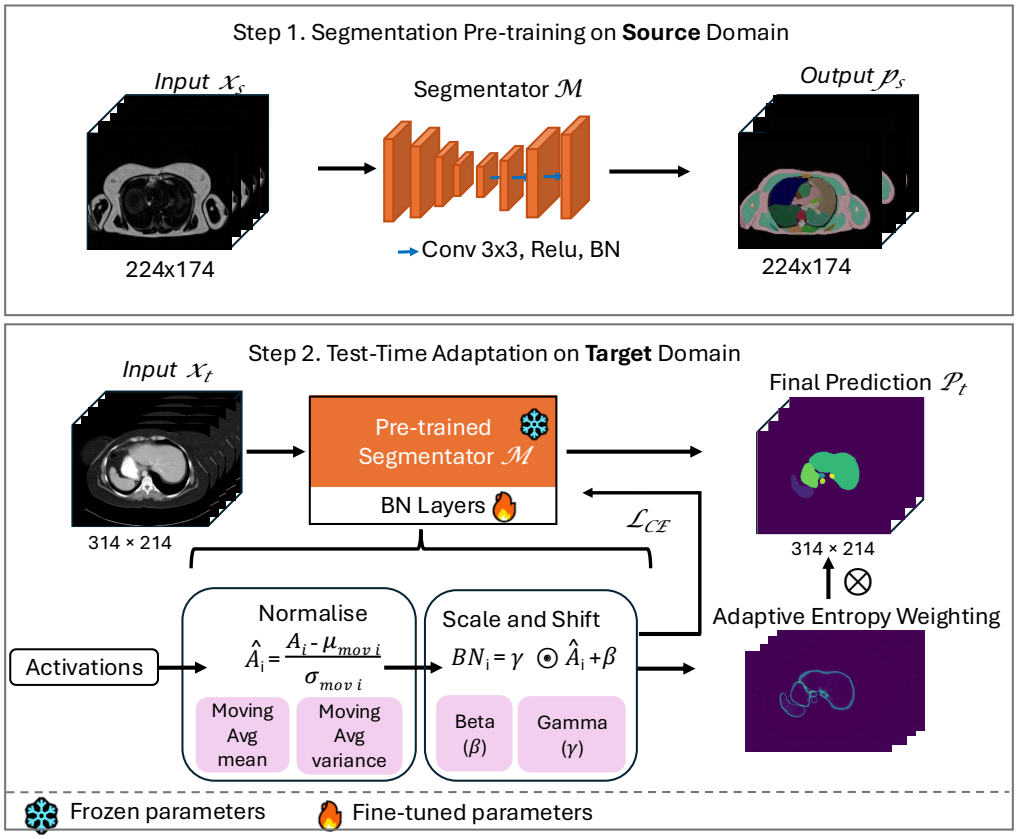


Figure 7.5: **Entropy Test-Time Adaptation for Image Segmentation.** We use a test-time entropy map to refine the batch norm layer of the network for robust segmentation output. This module is agnostic to the architecture of the deep neural network. Therefore, It can be used with any segmentation network to increase consistency and robustness, especially when trained with noisy labels.

the sample. Here, f_θ represents the segmentation network parameterized by θ. We define the entropy loss $\mathcal{L}_{\text{ent}}$ over the predicted probabilities:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} p_{i,c}\log p_{i,c} \tag{7.2}$$

where $p_{i,c}$ is the probability of class c at voxel i.

We update only the batch normalization parameters θ_{BN} while keeping the other network parameters θ_{fixed} frozen (see pipeline in Figure 7.5). The adaptation process involves minimizing the entropy loss with respect to θ_{BN}:

$$\theta^*_{\text{BN}} = \arg\min_{\theta_{\text{BN}}} \mathcal{L}_{\text{ent}}\left(f_{\theta_{\text{fixed}},\,\theta_{\text{BN}}}(\mathbf{x})\right) \tag{7.3}$$

This process adapts the model to the test sample by encouraging confident (low-entropy) predictions, thereby refining the segmentation output. The adaptation is efficient as it involves updating a small subset of parameters and can be performed online during inference. ETTA leverages the entropy of the network's predictions (Figure 7.6) to guide

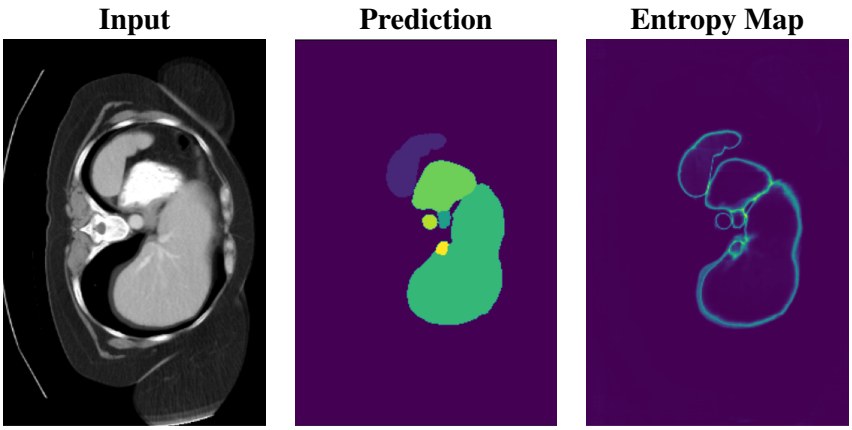


Figure 7.6: **Entropy Map Visualization.** We show from left to right an example of input, prediction and entropy map used in the Entropy Test-Time Adaptation on BTCV dataset [Fang and Yan 2020] that can be leveraged to refine the output. Brighter regions indicate higher entropy.

the adaptation, improving robustness to domain shifts and label noise.

7.4 Experiments

7.4.1 Evaluation Datasets and Metrics

Evaluation Datasets. We evaluate our model on multi-modal publicly available datasets i.e. AMOS [Ji et al. 2022] of 13 abdominal organs from 100 MRI scans split equally into train and test sets, BTCV (Beyond the Cranial Vault) abdomen CT dataset [Fang and Yan 2020] of 30 training and 20 testing subjects and 13 labelled organs, and BRATS [Baid et al. 2021; Menze et al. 2015; Bakas et al. 2017], the largest publicly available dataset for brain tumors 5,880 MRI scans and corresponding annotations.

Evaluation Metrics. We evaluate our model using the Dice Score and the Hausdorff Distance Metric, which are widely used in medical image segmentation [Ma et al. 2024; Karimi and Salcudean 2019]. The *Dice Score* measures the overlap between predicted and ground truth masks, while the *Hausdorff Distance* assesses the boundary discrepancy, providing a comprehensive evaluation of segmentation performance.

7.4.2 Baselines.

To evaluate our model, we compared it with a variety of established baselines across the BTCV, BRATS, and UKBOB benchmarks. For the BTCV dataset, baseline models included UNet [Ronneberger et al. 2015], SegResNet [Myronenko 2018], TransUNet [Chen et al. 2021], UNetr [Hatamizadeh et al. 2021], Swin-UNetr [Hatamizadeh et al. 2022], nn-UNet [Isensee et al. 2021], and AttentionUNet [Oktay et al. 2018]. We also evaluate the base TotalVibeSegmentator (TVS) [Graf et al. 2024] in zero-shot and finetuning settings. These models represent widely adopted architectures in medical image segmentation, offering a range of network designs from classic CNN-based approaches to transformer-based architectures. For the BTCV dataset, diffusion-based segmentation methods were also considered, including MedSegDiff [Wu et al. 2022] and MedSegDiff-V2 [Wu et al. 2024]. Together, these baselines provide a comprehensive foundation for evaluating our model's performance in 3D medical segmentation. On the BRATS2023 benchmark, additional baselines incorporated V-Net [Milletari et al. 2016], ResUNet++ [Jha et al. 2019], and nnFormer [Zhou et al. 2023], along with UNETR [Hatamizadeh et al. 2021] and Swin-UNetr [Hatamizadeh et al. 2022]. These models were selected to encompass a spectrum of segmentation methods specifically suited to brain tumor segmentation tasks.

Model	Dice Score	Hausdorff Distance
ResUNet++ [Jha et al. 2019]	0.876	9.431
MedFormer [Wang et al. 2024b]	0.881	8.822
nnUNet [Isensee et al. 2021]	0.915	6.442
UNetr [Hatamizadeh et al. 2021]	0.902	7.968
Swin-UNetr [Hatamizadeh et al. 2022]	**0.918**	**5.984**

Table 7.2: **UKBOB 3D Segmentation Benchmark.** We show results on test mean Dice Score (%) and mean Hausdorff Distance ($n = 72$ classes) of our proposed benchmark on UKBOB. Note how Swin-UNetr [Hatamizadeh et al. 2022] achieves the best results, resulting in our Swin-BOB foundation model.

7.4.3 Implementation Details

Pre-training. For pre-training on UKBiobank, we split the dataset into 80-10-10 for training, validation, and testing. The input is cropped to 96×96×96 voxels from the 3D MRI. In training, for data augmentation, scans are intensity scaled, with random flipping along the 3 axes, random foreground cropping, random rotation 90 degrees with probability 10%, and random intensity shift with an offset of 0.1. In validation, scans are intensity-scaled. We use a batch size of 8, AdamW optimizer, β_1 = 0.9, β_2 = 0.999, and cosine learning rate scheduler with a warm restart every 200 epochs. We start with the initial learning rate of 10^{-4} and decay of 10^{-5}. We train the model on 2 A6000 GPU for 3,000 epochs. We use binary Cross-Entropy and Dice Similarity Coefficient (DSC) as our loss function. The best model achieving the highest overall Dice score on the 72 classes is saved at validation.

Fine-tuning. The pre-trained Swin-BOB is used on BTCV, BRATS, and AMOS, keeping the same configuration as in training while reducing the warm-up scheduler to 50 epochs for a total of 500 epochs.

7.5 Results

7.5.1 Validating UKBOB Labels

Manual Verification. In Table 7.3 we validate the quality of the UKBOB labels and the organ quality control against manual spine annotation from [Bourigault et al. 2023] and our manually annotated 11 abdomen UKBOB organs (UKBOB-manual). Even without any filtration, UKBOB labels are precise, achieving Dice score of 0.811 and 0.873 on manual spines and UKBOB-manual respectively. We Also show that our SOLF filtering approach (when $\epsilon = 2$) increases Dice score by 0.056 compared to no filtering

Configuration	No Filter	IQR filter	SOLF Filter
Spine labels [Bourigault et al. 2023]	0.811	0.849	0.867
UKBOB-manual	0.873	0.877	0.891

Table 7.3: **Precision of the collected UKBOB compared to Manual labels.** We show the Dice Score of the collected UKBOB labels on subsets of manual labels on UKBOB for 200 spines [Bourigault et al. 2023] and on 300 manual abdominal labels we collect (UKBOB-manual). Even without any filtration, UKBOB labels are precise, improving in precision with our designed SOLF statistical filter.

on the spine labels and by 0.018 on the labelled abdomen organs. We also show standard inter-quartile range filtering (IQR) [Ji et al. 2022] for comparison.

Zero-Shot Evaluations. We also rely on the zero-shot segmentation performance of a model trained solely on those collected labels to be evaluated on a similar domain, namely the AMOS Abdomen MRI dataset [Ji et al. 2022] and on MRI on BTCV dataset [Fang and Yan 2020] in Table 7.5. AMOS shares 12 class labels with UKBOB while BTCV shares 11 class labels. In Table 7.5, we show 10 organs overlap between BTCV and AMOS where we omitted small organ i.e. duodenum as it has also been omitted in baseline papers. We combined the left and adrenal gland into one class named AG. We train Swin-UNetr [Hatamizadeh et al. 2022] from scratch on different filtration schemes of the UKBOB and run the evaluation on the test sets given by BTCV and AMOS (on the shared class labels) and report the mean Dice score and mean Haussdorff distance. We adjust the preprocessing (normalization to [0,1], and resizing) to ensure compatibility with the model's pretrained dataset. These results highlight the importance of the filtration we followed ensuring better quality labels.

7.5.2 Swin-BOB: A Foundation Model for 3D Medical Image Segmentation

We train Swin-UNetr [Hatamizadeh et al. 2022] on our filtered UKBOB for a foundation model (Swin-BOB) for 3D segmentation. We evaluate test performance on multiple downstream tasks including BRATS brain MRI benchmark [Baid et al. 2021] in Table 7.4 and BTCV Abdomen CT [Fang and Yan 2020] in Table 7.7 (examples in Figure 7.7). In both benchmarks, our Swin-BOB achieves state-of-the-art with up to 0.02 Dice score improvement and reduction of 2.4 in Mean Hausdorff Distance. We also establish a UKBOB benchmark with reported Dice score and Mean Hausdorff Distance of different networks in Table 7.2 to aid research in this direction.

Model	Dice Score	Hausdorff Distance
UNet[Ronneberger et al. 2015]	0.544	39.090
V-Net[Milletari et al. 2016]	0.842	10.891
ResUNet++[Jha et al. 2019]	0.784	22.249
AttentionUNet[Oktay et al. 2018]	0.798	20.048
nnFormer[Zhou et al. 2023]	0.812	10.070
UNETR[Hatamizadeh et al. 2021]	0.871	9.924
SegResNet[Myronenko 2018]	0.890	8.650
Total Vibe Seg.[Graf et al. 2024]	0.830	8.973
Swin-UNetr[Hatamizadeh et al. 2022]	0.886	9.016
Swin-BOB (ours)	**0.894**	**8.650**

Table 7.4: **BRATS 3D Segmentation Benchmark.** The proposed Swin-BOB model, pre-trained on UK Biobank organs and fine-tuned on BRATS2023 [Baid et al. 2021] archives state-of-the-art results on test mean Dice Score (%) and mean Hausdorff Distance ($n = 3$ classes). Baseline results are reported from the Swin-UNetr paper [Hatamizadeh et al. 2022].

7.5.3 Entropy Test Time Adaptation Results

In Table 7.6 we show the benefit of endowing different fine-tuned models with our proposed ETTA and show improvement on 3 different datasets' test performance for 3D segmentation using 3 different networks (including Swin-BOB). In all the 3 networks, we compare the ETTA against augmentation-based test-time adaptation baseline [Zhang et al. 2020]. This highlights the importance of ETTA in tackling the issue of domain shift in medical imaging especially when the training includes noisy labels, as in the case of the Swin-BOB model.

Dataset	Config.	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	IVC	AG	Aorta	Mean
AMOS	TVS	0.823	0.697	0.814	0.782	0.786	0.802	0.759	0.738	0.879	0.881	0.796
	no filter	0.908	0.931	0.942	0.657	0.658	0.958	0.822	0.874	0.529	0.906	0.818
	+ vol. filter	0.910	0.940	0.951	0.658	0.667	0.966	0.832	0.882	0.621	0.918	0.832
	+ SOLF filter	0.919	0.943	0.962	0.664	0.672	0.969	0.838	0.882	0.631	0.924	0.840
BTCV	TVS	0.848	0.721	0.801	0.785	0.797	0.795	0.737	0.715	0.861	0.862	0.792
	no filter	0.883	0.884	0.932	0.795	0.790	0.946	0.885	0.871	0.784	0.799	0.856
	+ vol. filter	0.881	0.870	0.939	0.812	0.801	0.925	0.873	0.858	0.781	0.852	0.875
	+ SOLF filter	0.891	0.890	0.949	0.823	0.881	0.897	0.899	0.891	0.824	0.871	0.882

Table 7.5: **Detailed Zero-shot 3D Segmentation Performance.** We show Zero-shot Test Dice Score of Swin-BOB on AMOS external MRI data and CT (BTCV) for same organ classes. We show 10 organs that overlap between BTCV and AMOS where we combined left and right adrenal gland into one class named `AG` while inferior vena cava is briefed as `IVC`. TotalVibe Segmentator model (TVS) [Graf et al. 2024] results are shown for reference, while the Swin-BOB model is trained on complete UKBOB without filtration, with normalized volume statistical filter, and with full SOLF filter respectively. Note the significant benefit of filtering the collected UKBOB labels.

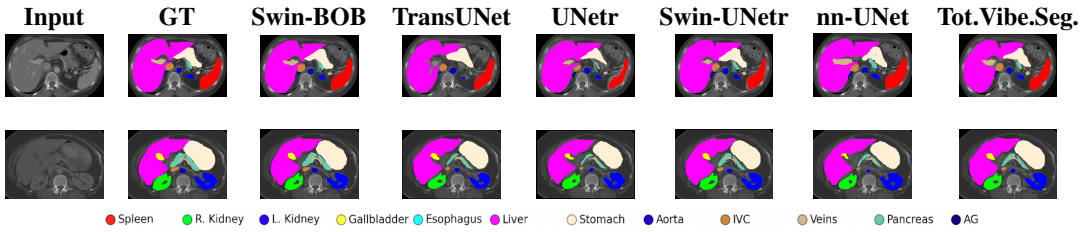


Figure 7.7: **Qualitative Results on BTCV.** We show comparisons of 3D segmentation on the abdomen BTCV dataset with 12 organ labels [Fang and Yan 2020]. Note the significant improvement of our Swin-BOB especially on *Stomach* in the first row and *Pancreas* in the second row

Configuration	BTCV		AMOS		BRATS	
	Mean Dice	Mean HD	Mean Dice	Mean HD	Mean Dice	Mean HD
nn-UNet [Isensee et al. 2021]	0.804	12.141	0.795	9.623	0.812	9.787
nn-UNet + TTA [Zhang et al. 2020]	0.811	10.901	0.830	8.465	0.832	8.327
nn-UNet + ETTA (ours)	0.831	8.652	0.826	7.683	0.848	7.874
Swin-UNetr [Hatamizadeh et al. 2022]	0.872	8.517	0.822	8.390	0.885	8.929
Swin-UNetr + TTA	0.870	8.280	0.839	7.726	0.880	8.654
Swin-UNetr + ETTA (ours)	0.886	7.221	0.858	5.812	0.894	7.463
Swin-BOB (ours)	0.883	8.261	0.847	8.105	0.882	8.624
Swin-BOB + TTA	0.883	7.901	0.857	5.651	0.887	7.712
Swin-BOB + ETTA(ours)	**0.892**	**7.381**	**0.864**	**7.191**	**0.894**	**7.130**

Table 7.6: **Effect of Entropy Test-Time Adaptation (ETTA).** We demonstrate that the proposed ETTA enhances the performance of fine-tuned models on the BTCV [Fang and Yan 2020], BRATS [Baid et al. 2021], and AMOS [Ji et al. 2022] datasets. The best results are achieved by fine-tuning our baseline Swin-BOB—pre-trained on the UKBOB dataset. Our ETTA consistently improves performance across various networks and downstream tasks, outperforming the standard TTA baseline [Zhang et al. 2020].

7.5.4 Analysis and Insights

Filtration Ablation Study. We study the effect of the filtration threshold ϵ in the SOLF filter (Eq (7.1) of the three features) on the zero-shot generalization of the models trained on the filtered subsets of UKBOB. For ϵ = 0, 1, 2, 3, 4, and 5, the performance Dice score on BTCV is 0.792, 0.884, 0.892, 0.766, and 0.745, respectively. We also ablate the features used in the SOLF filter. In Table 7.5, when only the normalized volume is used in the SOLF filter(no Sphericity or Eccentricity), the quality of the filtration degrades considerabley, highlighting the importance of each aspect of the SOLF filter to clean the labels.

Filtering Out Patients Abnormalities. One concern of automatic filtration in Section 7.3.2 is that it might filter out some natural abnormalities or pathologies in the patients, mistaken as wrong labels. We visualize some of these filtered-out labels in Figure 7.9 and show that indeed lack quality labels rather than the patients have obvious abnormalities. The combination of normalized volume, sphericity and eccentricity makes the filtration mostly about the quality of the labels rather than filtering out patients with

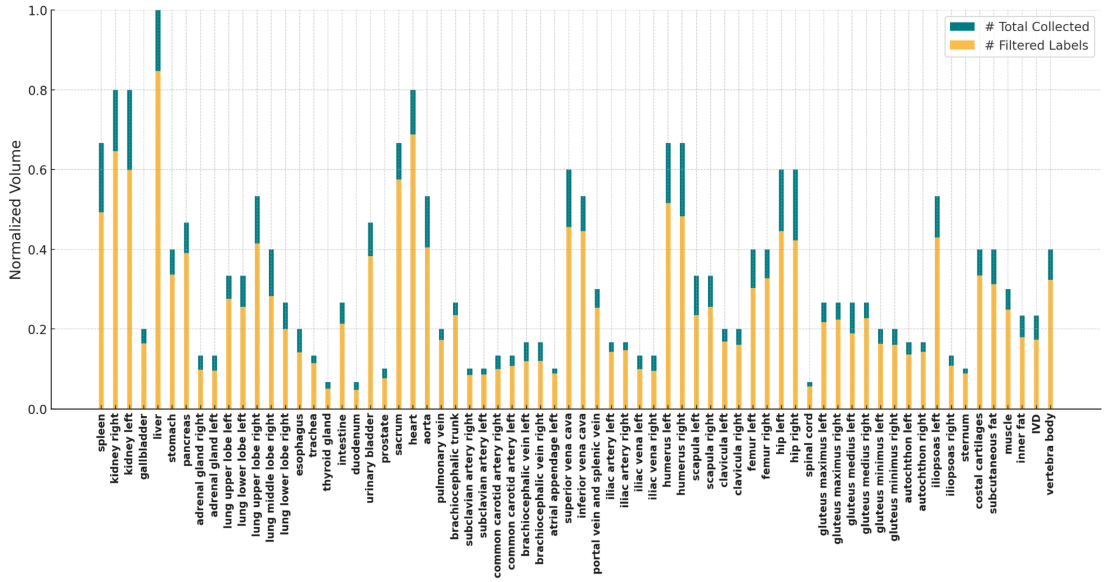


Figure 7.8: **UKBOB Distribution of Labels with our Filtration.** We show the distribution of mean normalized volumes of 72 labels before and after SOLF filtration. More examples and classes details are available in *Appendix*.

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Avg.
TransUNet [Chen et al. 2021]	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
UNetr [Hatamizadeh et al. 2021]	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNetr [Hatamizadeh et al. 2022]	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	0.869
nnUNet [Isensee et al. 2021]	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
Total Vibe Seg. [Graf et al. 2024]	0.948	0.914	0.917	0.736	0.741	0.954	0.859	0.881	0.794	0.752	0.718	0.699	0.826
MedSegDiff [Wu et al. 2022]	0.973	0.930	0.955	0.812	0.815	0.973	0.924	0.907	0.868	0.825	0.788	0.779	0.879
Swin-BOB (ours)	0.979	0.951	0.967	0.815	0.792	0.984	0.937	0.909	0.870	0.882	0.832	0.796	**0.892**
MedSegDiff-V2 (ens.)[Wu et al. 2024]	0.978	0.941	0.963	0.848	0.818	0.985	0.940	0.928	0.869	0.823	0.831	0.817	0.895
Swin-BOB (ens.)	0.981	0.958	0.971	0.817	0.796	0.988	0.942	0.912	0.874	0.886	0.836	0.799	**0.897**

Table 7.7: **3D Segmentation Performance on the BTCV Benchmark.** We evaluate our approach on the 3D segmentation task of the BTCV dataset using the Dice score. For a fair comparison, we also report 10-fold ensembling results (denoted as *ens.*) as presented in MedSegDiff-V2 [Wu et al. 2024].

abnormality. Figure 7.8 shows the distribution of organs normalized volumes in UKBOB before and after filtration.

Scaling-Law of 3D Medical Segmentation. We study the impact of the scale of training data of UKBOB on the downstream 3D segmentation performance, to justify the large scale UKBOB. We independently trained the Swin-UNetr network on subsets of UKBOB (i.e. 0%, 20%, 40%, 60%, and 80% and show in Figure 7.10 the Average test Dice Score for both BTCV [Fang and Yan 2020] and BRATS [Baid et al. 2021]. We also show the t-sne visualization of the features in Figure 7.11 illustrating the quality of the features.

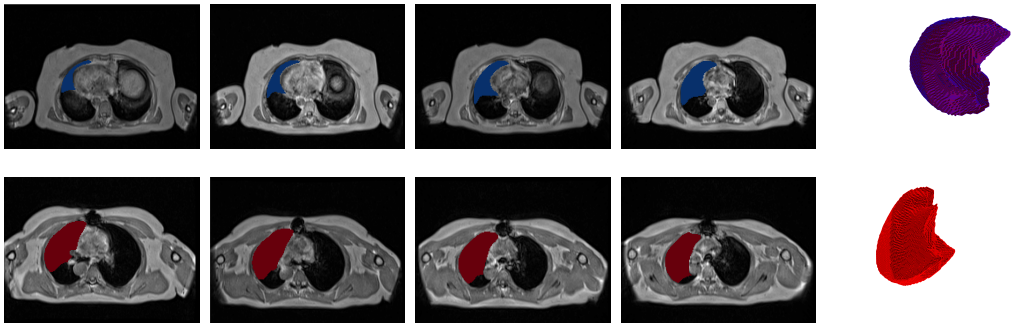


Figure 7.9: **Filtration of Inaccurate Labels.** *top* are manual labels of upper left lung overlaid on scans for several slice indices. *bottom* are filtered out labels (red) of the upper left lung overlaid on scans with corresponding slice indices. The filtered out lung is incomplete and erroneous.

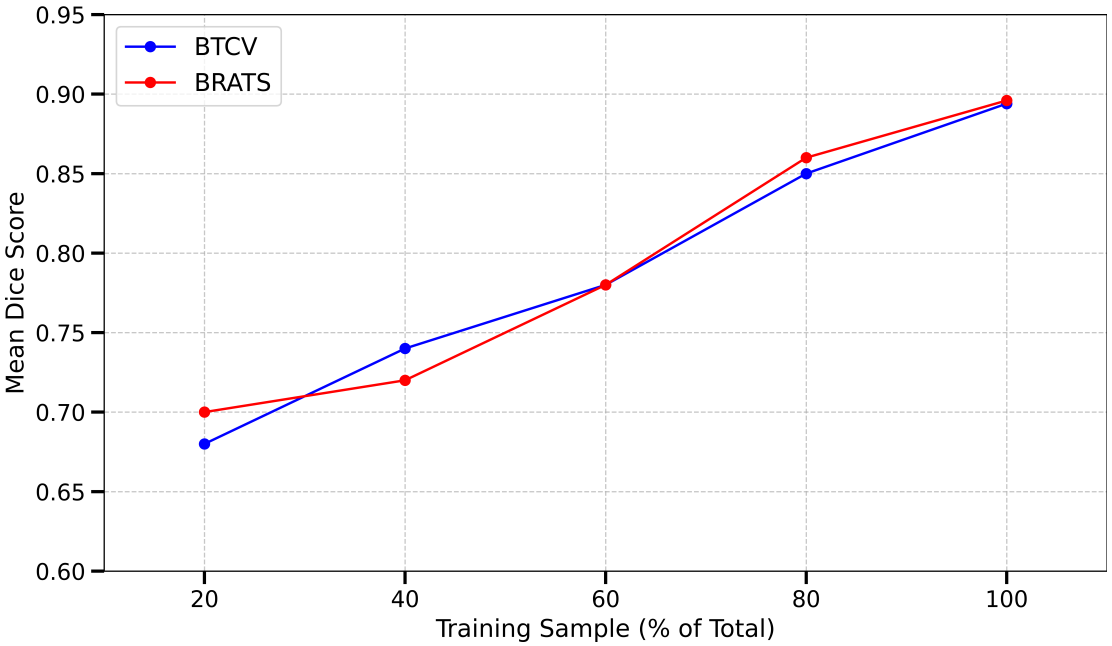


Figure 7.10: **Effect of UKBOB Pre-Training Dataset Size on Downstream Segmentation Performance.** We observe a consistent increase in test Dice Score for both BRATS [Baid et al. 2021] and BTCV [Fang and Yan 2020] when doubling the size of pre-training Swin-BOB on UKBOB, acting as a foundation model.

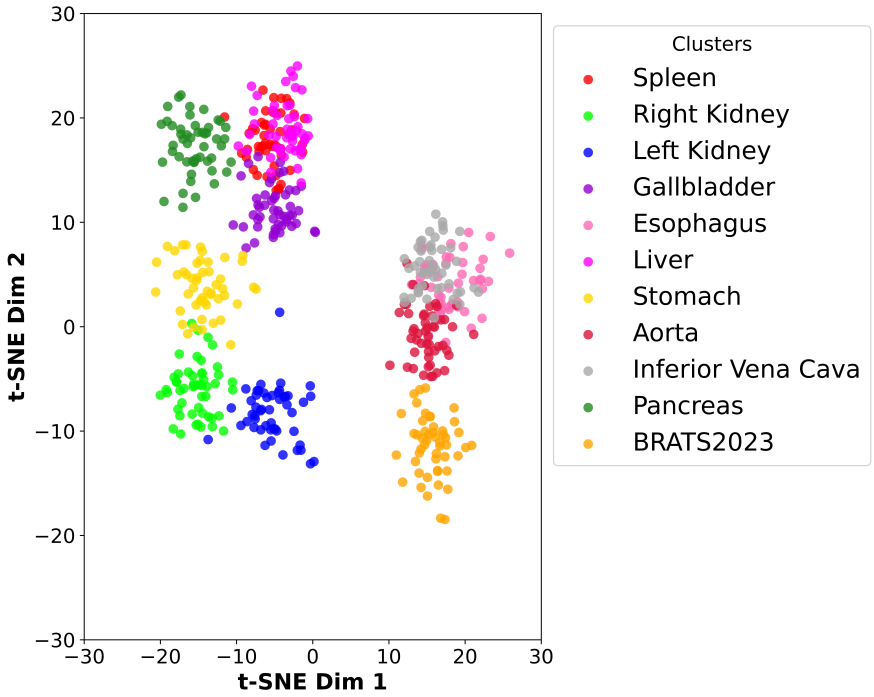


Figure 7.11: **Distribution of Feature Embeddings on BTCV organs and BRATS23**. Each category is represented with a unique color. We reduce features embeddings to 2D for each class using t-sne [Van der Maaten and Hinton 2008]. The low dispersion of the clusters between each other indicates that the features of different classes probably share similar patterns and this explains the beneficial effect of large pre-training.

7.6 Conclusions and Future Works

In this work, we introduced the UKBOB dataset, the largest labeled medical imaging dataset to date, comprising $51,761$ 3D MRI scans and over 1.37 billion 2D segmentation masks covering 72 organs. Our models trained on UKBOB demonstrate strong zero-shot generalization to other medical imaging datasets and achieve state-of-the-art performance on several benchmarks in 3D medical image segmentation.

Limitations and Future Works. While UKBOB significantly expands the availability of large-scale labeled data for medical imaging, it is limited to neck-to-knee MRI scans and may not encompass the full diversity of imaging modalities and anatomical regions. Despite our filtration process, the automatic labeling may still introduce residual label noise that could impact model training. Future work includes extending the dataset to cover additional imaging modalities such as CT scans and incorporating more anatomical regions. Additionally, exploring advanced adaptation techniques and integrating clinical metadata could enhance model robustness and applicability across diverse clinical settings.

Acknowledgments. This work was supported by the Centre for Doctoral Training in Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research (SABS: R3), University of Oxford (EP/S024093/1), and by the EPSRC Programme Grant Visual AI (EP/T025872/1). We are also grateful for the support from the Novartis-BDI Collaboration for AI in Medicine. Part of the support is also coming from KAUST Ibn Rushd Postdoc Fellowship program.

A Detailed Setup

A.1 Datasets

We conducted our experiments on four primary datasets:

1. **UK Biobank** A more comprehensive dataset of 51,761 full-body MRIs from more than 50,000 volunteers [Sudlow et al. 2015], capturing diverse physiological attributes across a broad demographic spectrum. UK Biobank MRIs are resampled to be isotropic and cropped to a consistent resolution ($501 \times 160 \times 224$).

2. **BRATS** The largest public dataset of brain tumours consisting of 5,880 MRI scans from 1,470 brain diffuse glioma patients, and corresponding annotations of tumours [Baid et al. 2021; Menze et al. 2015; Bakas et al. 2017]. All scans were skull-stripped and resampled to 1 mm isotropic resolution. All images have resolution $240 \times 240 \times 155$. Tumours are annotated by expert clinicians for three classes: Whole Tumour (WT), Tumour Core (TC), and Enhanced Tumour Core (ET).

3. **BTCV** (Beyond the Cranial Vault) abdomen dataset [Fang and Yan 2020]. This dataset involves 30 training and 20 testing subjects and 13 labelled organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland and left adrenal gland. We combine the left and right adrenal gland into one. Scans are resampled to consistent resolution ($224 \times 224 \times 85$) and intensity scaled in the range [-175,250] Hounsfield Units (HU).

4. **AMOS** Abdomen MRI [Ji et al. 2022] from the MICCAI AMOS Challenge, which consists of segmentation of abdominal organs from 100 MRI scans split equally into train and test sets. The organs include the liver, spleen, pancreas, kidneys, stomach, gallbladder, esophagus, aorta, inferior vena cava, adrenal glands, and duodenum. Scans are resampled to consistent resolution ($256 \times 256 \times 125$) and scans normalised for intensity channel wise in the range [0,1].

A.2 Evaluation Metrics

- **Dice Score** The Dice Score, or Dice Coefficient, is a statistical measure used to assess the similarity between two samples. It is widely utilized in medical image analysis due to its sensitivity to variations in object size. The Dice Score is

calculated by doubling the area of overlap between the predicted and ground truth segmentations and dividing by the total area of both. The formula is:

$$\text{Dice} = \frac{2 \times \text{Area}(S_{\text{pred}} \cap S_{\text{gt}})}{\text{Area}(S_{\text{pred}}) + \text{Area}(S_{\text{gt}})}$$

This metric ranges from 0 to 1, with a value of 1 indicating perfect agreement between the prediction and the ground truth. The Dice Score is particularly robust against variations in the size of the segmented objects, making it extremely useful in medical applications where such variability is common.

Both IoU and Dice Score offer comprehensive insights into model accuracy, with the Dice Score being especially effective in scenarios involving significant variations in object size.

- **Hausdorff Distance** The Hausdorff Distance is a metric used to measure the extent of discrepancy between two sets of points, often applied to evaluate the accuracy of object boundaries in image segmentation tasks. It is particularly useful for quantifying the worst-case scenario of the distance between the predicted segmentation boundary and the ground truth boundary.

  The Hausdorff Distance calculates the greatest distance from a point in one set to the closest point in the other set. In image segmentation, this involves finding the largest distance from any point on the predicted boundary to the nearest point on the ground truth boundary, and vice versa. The mathematical definition is:

$$\text{HD} = \max\left\{\sup_{p \in P} \inf_{q \in Q} d(p, q), \sup_{q \in Q} \inf_{p \in P} d(p, q)\right\}$$

  where P and Q are the sets of boundary points of the predicted segmentation and the ground truth segmentation, respectively, and $d(p, q)$ represents the Euclidean distance between points p and q.

A.3 Segmentation Details

We perform a series of experiments to determine the best segmentation model on UK-BOB using state-of-the-art multi-resolution CNN (UNet [Ronneberger et al. 2015], SegResNet [Myronenko 2018], nn-UNet [Isensee et al. 2021]) and transformer-based networks (TransUNet [Chen et al. 2021], UNetr [Hatamizadeh et al. 2021], Swin-UNetr [Hatamizadeh et al. 2022]). We report segmentation performance in Table B6 where

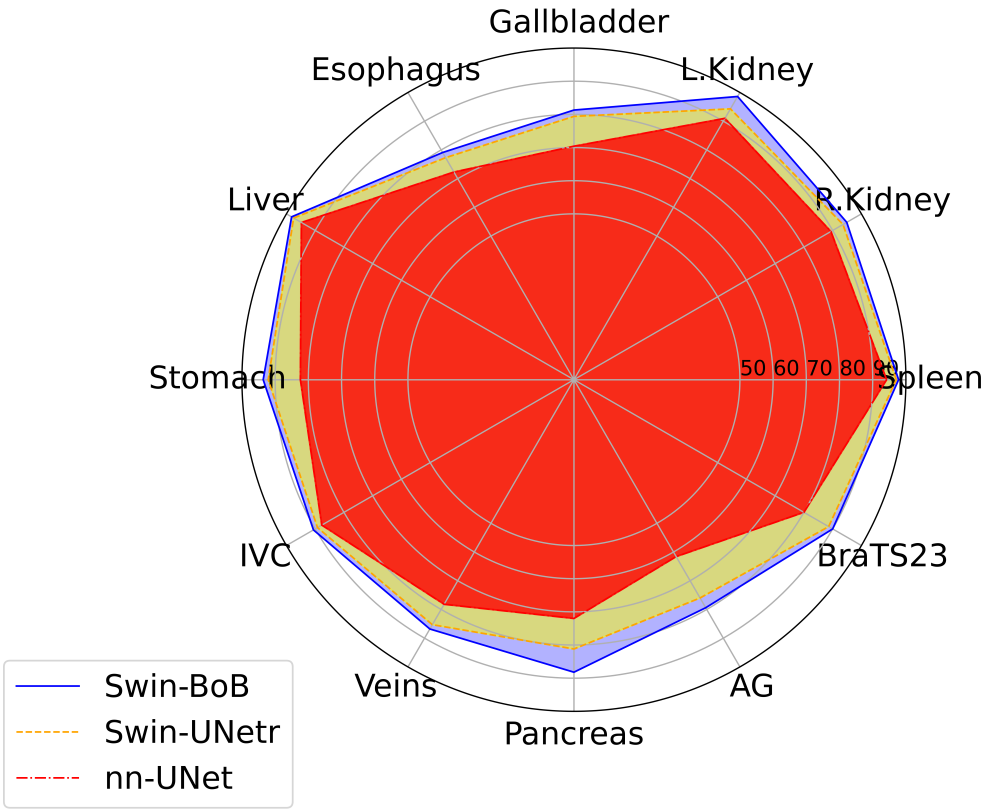


Figure A1: **Per-Class Performance Comparison with Specialized Segmentation models.** We compare the Dice Score performance of our Swin-BOB model and baselines Swin-UNetr [Hatamizadeh et al. 2022] and nn-UNet [Isensee et al. 2021] on abdominal organ segmentation (BTCV) and brain tumour segmentation (BRATS).

Swin-UNetr outperforms baselines by a margin, followed by nn-UNet. We show visual examples of the 72 class labels in UKBOB in Figure B3 and Figure B4.

We also show detailed baseline comparison for BTCV and AMOS in Table A1 and Table A2 respectively. We provide radar plot in Figure A1 that summarizes the performance of our segmentation model Swin-BOB compared to baseline segmentation models on different classes from BTCV and BRATS23 class average.

Model	Mean Dice Score	Mean Hausdorff Distance
UNet[Ronneberger et al. 2015]	0.782	8.374
SegResNet[Myronenko 2018]	0.794	7.912
TransUNet[Chen et al. 2021]	0.838	6.258
UNetr[Hatamizadeh et al. 2021]	0.856	4.317
Swin-Unetr[Hatamizadeh et al. 2022]	0.869	3.801
nn-UNet[Isensee et al. 2021]	0.802	6.782
AttentionUNet[Oktay et al. 2018]	0.816	5.848

Table A1: **Comparison of segmentation model performance on BTCV (n = 12 classes).**

We show visual comparison on BRATS (Figure A2) of our model segmentation relative to ground-truth.

Model	Mean Dice Score
TransBTS[Wang et al. 2021c]	0.792
UNETR[Hatamizadeh et al. 2021]	0.762
nnFormer[Zhou et al. 2023]	0.790
SwinUNETR[Hatamizadeh et al. 2022]	0.880
3D UX-Net[Lee et al. 2023]	0.900

Table A2: **Comparison of Segmentation Models for AMOS Segmentation (n = 14 classes).**

Model	Mean Dice Score	Mean Hausdorff Distance
$\epsilon = 3$	0.891	7.126
$\epsilon = 2$	0.884	7.528
$\epsilon = 1$	0.792	8.247
$\epsilon = 4$	0.766	8.594
$\epsilon = 5$	0.745	8.972

Table A3: **Effect of Filtration Threshold on Segmentation Performance on manual annotated set of abdomen organs (300) from UK Biobank.** The 11 abdomen organs and bones that have been manually annotated represent the overlap organs with BTCV [Fang and Yan 2020] and UK Biobank [Graf et al. 2024].

Dataset	Mean Dice Score	Hausdorff Distance
AMOS	0.831	7.647
BTCV	0.837	5.138

Table A4: **Zero-shot performance on external datasets.**

Configuration	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	IVC	AG	Aorta
AMOS	0.9084	0.9311	0.9421	0.6516	0.6582	0.9581	0.8216	0.8740	0.5292	0.9062
AMOS + filtering	0.9102	0.9397	0.9508	0.6582	0.6673	0.9662	0.8315	0.8824	0.6209	0.9183
BTCV	0.883	0.884	0.932	0.795	0.790	0.946	0.885	0.871	0.784	0.799
BTCV + filtering	0.889	0.889	0.941	0.813	0.825	0.949	0.893	0.883	0.799	0.869

Table A5: **Zero-shot 3D Segmentation Performance of Swin-BOB on AMOS external MRI data and CT (BTCV) for same organ classes**.

A.4 Threshold Selection

Full ablation experiments for threshold selection is available in Table A3. Results on impact of filtration on BTCV and AMOS are reported in Table A5. We therefore ensure high-quality labels by removing outliers adequately.

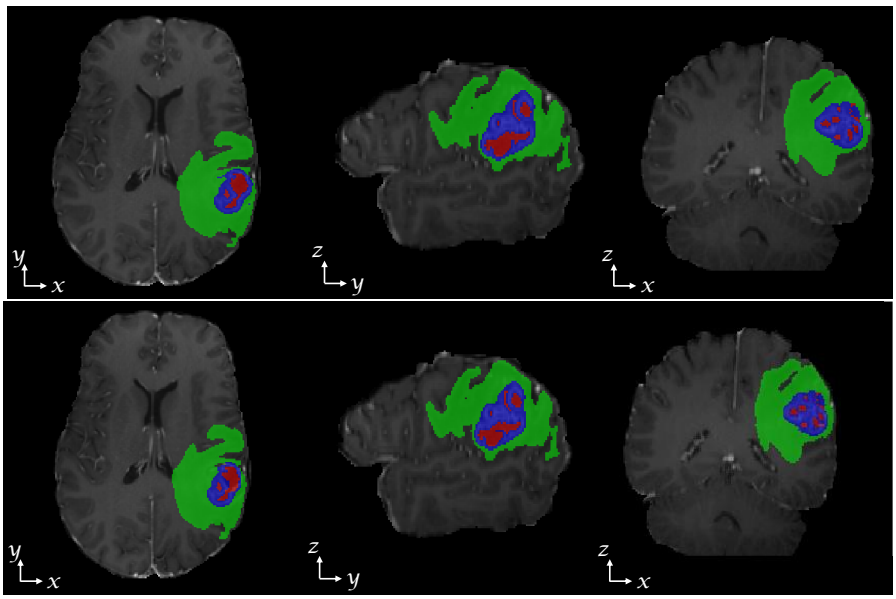


Figure A2: **Qualitative Performance on BRATS.** We show the ground-truth *top* and output *bottom* of our pre-trained Swin-BOB model for 3D segmentation on the brain tumour BRATS dataset with 3 tumour class labels [Baid et al. 2021].

A.5 Zero-Shot Generalization

Our zero-shot evaluation on the AMOS and BTCV datasets highlights the robustness of filtered labels. Metrics are detailed in Table A4.

A.6 Residual Label Noise

While filtration reduces label noise, some false positives persist. To further improve the quality of the segmentation, we could incorporate human-in-the-loop approaches that turned efficient as shown in [Graf et al. 2024; Bourigault et al. 2023].

A.7 Filtering Out Patients Abnormalities

One concern of automatic filtration is that it might filter out some natural abnormalities or pathologies in the patients, mistaken as wrong labels. We visualize some of these filtered-out labels in Figure 9 (main paper) and show that indeed lack quality labels rather than the patients have obvious abnormalities. To quantify this behavior, we measure the 50-sample average LPIPS distance (the lower the more similar) between any two 3D mid-abdominal slices from full UKBOB (0.315), between filtered/filtered-out samples (0.329), between filtered/filtered samples (0.303), and between filtered-out/filtered-out samples (0.339). This shows that all the distances are almost identical, indicating mostly homogeneous organs in the dataset partitioning and hence the filtration is mostly about the quality of the labels rather than filtering out patients with abnormality.

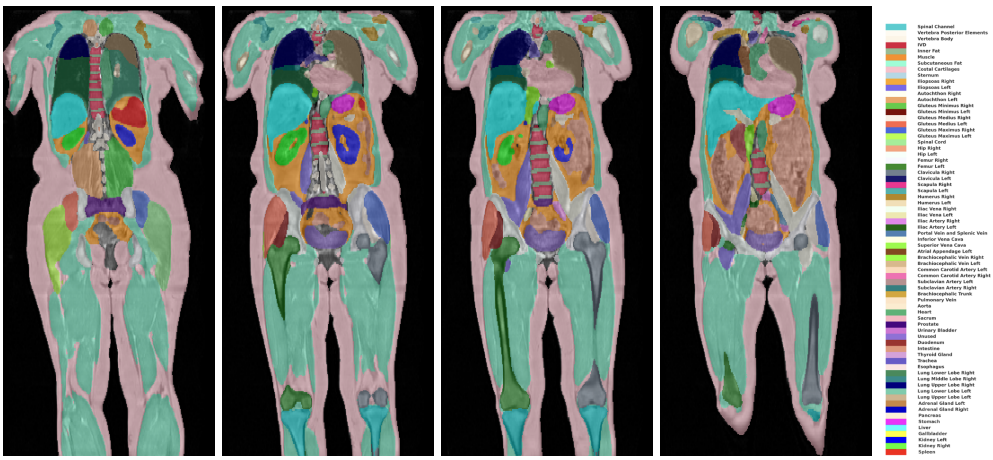


Figure B3: **Visualisation of UKBOB Segmentation Coronal Plane.** We show an example of 3D MRI from UKBOB for on coronal plane.

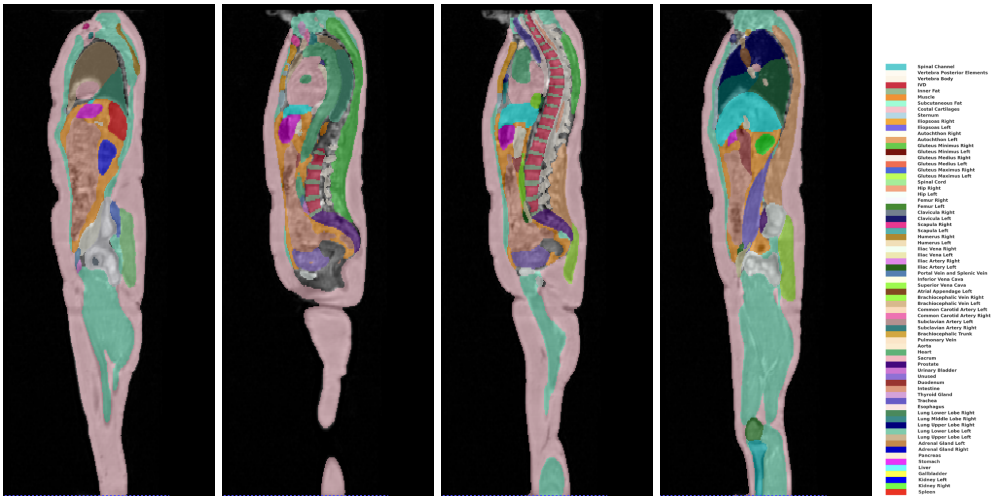


Figure B4: **Visualisation of UKBOB Segmentation Sagittal Plane.** We show an example of 3D MRI from UKBOB for on sagittal plane.

B Entropy Test-Time Adaptation (ETTA)

B.1 Algorithm Details

In this section, we detail the algorithmic process for our test-time adaptation (ETTA). It works by refining predictions minimizing entropy:

$$L_{\text{ent}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} p_{i,c}\log p_{i,c}.$$

During test time, only batch normalization parameters are fine-tuned while keeping other parameters fixed. The method is simple, and efficient computationally since it does not require retraining the full model.

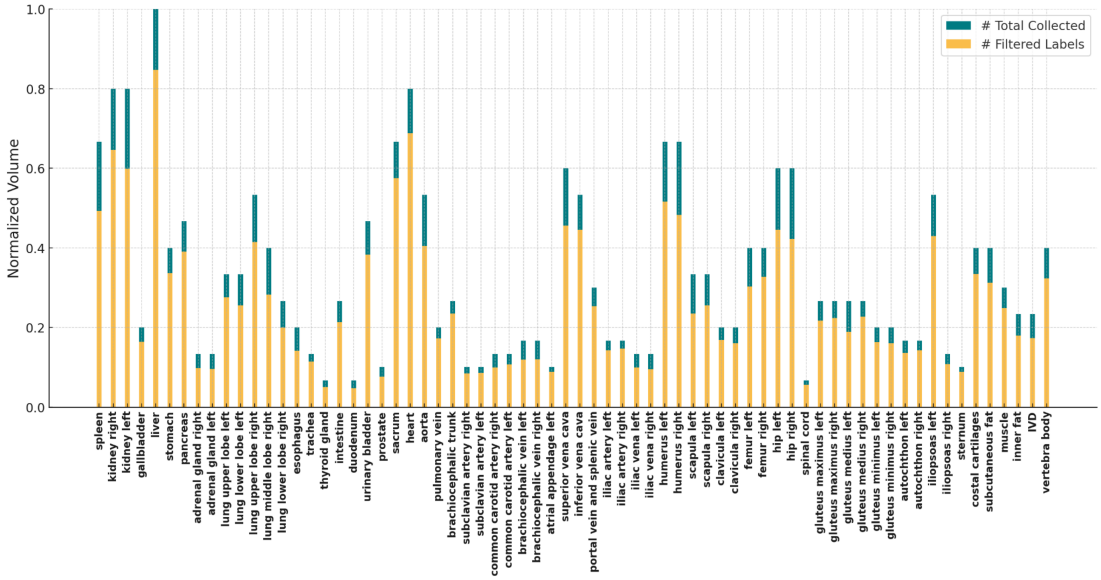


Figure B5: **UKBOB Distribution of Labels with our Filtration.** We show the distribution mean normalised volumes of 72 labels before and after filtration.

Model	ResUNet	UNetr	nnUNet	Swin-UNetr	MedFormer
spleen	0.91	0.92	0.94	0.94	0.93
kidney right	0.87	0.89	0.91	0.92	0.90
kidney left	0.88	0.90	0.92	0.93	0.91
gallbladder	0.82	0.84	0.85	0.85	0.84
liver	0.94	0.96	0.97	0.96	0.96
stomach	0.88	0.89	0.90	0.91	0.89
pancreas	0.85	0.87	0.89	0.90	0.88
adrenal gland right	0.81	0.83	0.84	0.86	0.84
adrenal gland left	0.81	0.83	0.84	0.86	0.83
lung upper lobe left	0.93	0.94	0.96	0.96	0.95
lung lower lobe left	0.94	0.95	0.96	0.96	0.94
lung upper lobe right	0.94	0.95	0.96	0.96	0.94
lung middle lobe right	0.93	0.94	0.96	0.96	0.95
lung lower lobe right	0.93	0.95	0.96	0.96	0,95
esophagus	0.86	0.88	0.9	0.91	0.89
trachea	0.87	0.89	0.92	0.92	0.91
thyroid gland	0.74	0.75	0.76	0.77	0.75
intestine	0.87	0.91	0.93	0.92	0.91
duodenum	0.81	0.84	0.86	0.87	0.85
urinary bladder	0.89	0.93	0.95	0.96	0.94
prostate	0.91	0.92	0.94	0.94	0.94
sacrum	0.91	0.92	0.96	0.95	0.04
heart	0.92	0.96	0.97	0.97	0.96
aorta	0.91	0.93	0.95	0.94	0.93
pulmonary vein	0.87	0.89	0.91	0.92	0.91
brachiocephalic trunk	0.83	0.86	0.88	0.89	0.88
subclavian artery right	0.81	0.85	0.86	0.88	0.86
subclavian artery left	0.81	0.85	0.86	0.88	0.86
common carotid artery right	0.81	0.83	0.84	0.86	0.86
common carotid artery left	0.81	0.83	0.85	0.87	0.85
brachiocephalic vein left	0.82	0.85	0.88	0.89	0.85
brachiocephalic vein right	0.83	0.85	0.87	0.88	0.85
atrial appendage left	0.79	0.82	0.84	0.84	0.83
superior vena cava	0.89	0.91	0.93	0.93	0.91
inferior vena cava	0.89	0.90	0.92	0.92	0.90
portal vein and splenic vein	0.76	0.79	0.82	0.82	0.80
iliac artery left	0.83	0.85	0.87	0.87	0.85
iliac artery right	0.82	0.84	0.86	0.86	0.84
iliac vena left	0.85	0.88	0.91	0.91	0.89
iliac vena right	0.85	0.88	0.90	0.90	0.88
humerus left	0.90	0.93	0.94	0.93	0.93
humerus right	0.90	0.93	0.94	0.94	0.93
scapula left	0.86	0.89	0.91	0.91	0.89
scapula right	0.88	0.89	0.91	0.91	0.89
clavicula left	0.86	0.88	0.90	0.90	0.88
clavicula right	0.86	0.88	0.90	0.90	0.88
femur left	0.91	0.94	0.97	0.96	0.95
femur right	0.90	0.93	0.96	0.95	0.93
hip left	0.92	0.95	0.97	0.98	0.96
hip right	0.91	0.94	0.96	0.97	0.95
spinal cord	0.85	0.87	0.88	0.90	0.88
gluteus maximus left	0.93	0.95	0.98	0.98	0.95
gluteus maximus right	0.93	0.95	0.98	0.98	0.95
gluteus medius left	0.94	0.97	0.98	0.98	0.97
gluteus medius right	0.94	0.97	0.97	0.97	0.97
gluteus minimus left	0.94	0.97	0.94	0.95	0.97
gluteus minimus right	0.94	0.97	0.94	0.95	0.97
autochthon left	0.94	0.96	0.97	0.97	0.96
autochthon right	0.94	0.96	0.97	0.97	0.96
iliopsoas left	0.92	0.94	0.96	0.96	0.95
iliopsoas right	0.91	0.93	0.96	0.96	0.95
sternum	0.86	0.88	0.92	0.92	0.89
costal cartilages	0.85	0.87	0.90	0.91	0.88
subcutaneous fat	0.89	0.92	0.95	0.96	0.93
muscle	0.91	0.93	0.96	0.97	0.94
inner fat	0.86	0.88	0.90	0.91	0.89
IVD	0.86	0.88	0.90	0.91	0.88
vertebra body	0.89	0.92	0.94	0.94	0.93
vertebra posterior elements	0.82	0.84	0.86	0.88	0.85
spinal channel	0.87	0.89	0.91	0.91	0.89
bone other	0.82	0.84	0.86	0.87	0.84

Table B6: **3D Segmentation Performance on UK Biobank dataset**. We compare our UKBOB on 3D medical segmentation task on the UK Biobank test set (n=10,353) with 5-fold cross validation compared to other methods using the average Dice score and average Haussdorff Distance (HD) per class as metric. Standard deviations are shown next to the mean Dice Score and HD values.

C Dataset Access and Code for Reproducibility

The dataset and pre-trained Swin-BOB models will be made available publicly via UK Biobank. Documentation for reproducing experiments is provided in supplementary materials. Code is available at https://github.com/EmmanuelleB985/UK_BOB/tree/main

Chapter 8

Summary and Extensions

In this work, we developed automated methods for continuous measurement of scoliosis handling two common challenges in the medical field namely domain gap and scarcity of human labels.

8.1 2D Scoliosis Measurement on DXA

Principal findings. We showed that an end-to-end, label-efficient pipeline can recover reliable maximum angle of the spine from low-dose DXA with limited manual annotations. A model trained on ALSPAC DXA scans without finetuning gave a Pearson's correlation of 0.89 on the UK Biobank DXA test set. We find that we gain a performance of 0.03 on the model trained on the UK Biobank with ALSPAC pseudo-labels. Iterative self-training boosted segmentation Dice from 0.79 to 0.92, and angle error fell to $1.8^\circ \pm 1.3^\circ$ versus expert measurements. These gains confirm that pseudo-labelling and bootstrapped curation are viable when dense annotation is infeasible.

Strengths and clinical relevance. DXA machines are ubiquitous in osteoporosis screening, so repurposing them for scoliosis monitoring offers immediate reach. Our spline-based angle computation mimics radiologist workflow, facilitating acceptance and easing regulatory review. Moreover, the 10s runtime on a CPU aligns with outpatient throughput constraints.

Limitations. (1) *Projection-specific bias.* DXA employs fan-beam geometry, and minor gantry tilt can distort apparent curvature. (2) *Postural variability.* Supine DXA

may underestimate deformity compared with standing radiographs. (3) *Domain drift.* We evaluated on GE Lunar scanners only while Hologic systems differ in energy spectrum and noise texture. (4) *Resolution heterogeneity.* Variable DXA resolution across manufacturers (ranging from 0.6mm to 1.2mm pixel spacing) and the fundamental difference from volumetric imaging limits cross-modal generalization, particularly when our models encounter older equipment with degraded detector sensitivity or newer ultra-high-resolution systems outside our training distribution.

Literature Context and Clinical Impact. This work directly addresses the annotation scarcity challenge identified by [Zhu 2005] and [Baur et al. 2017], surpassing existing semi-supervised methods. While [Yang et al. 2017] achieved state-of-the-art performance using 50% of training data, our iterative self-training requires even less initial annotation. Unlike MedAL [Smailagic et al. 2018] which achieved 80% accuracy with 425 labeled images, we reached 92% Dice through progressive pseudo-label refinement. This presents a fundamentally different approach from both active learning and traditional SSL reviewed in the literature. Clinically, this transforms the diagnostic pathway by enabling scoliosis screening during routine osteoporosis assessment, potentially capturing the 68% of adult scoliosis cases that are currently undiagnosed. The integration requires no additional imaging, addressing the cost barriers highlighted by [Cheplygina et al. 2019].

Future work. Domain-adversarial fine-tuning or test-time self-recalibration as explored later in Chapter 7 could address vendor heterogeneity. A paired DXA-radiograph study would quantify posture bias directly. Finally, incorporating vertebra-level uncertainty estimates (e.g. via Monte-Carlo dropout) could flag scans needing manual review. Critically, expanding our training cohort beyond the UK Biobank's predominantly white British population (94% European ancestry) to include diverse ethnic groups would improve generalization.

8.2 3D Shape Analysis of Scoliosis

Principal findings. Extending to volumetric MRI uncovered rotational and lordotic features invisible in 2D. The 3D spine mesh explained more variance in clinical severity than 2D angles alone and clustered patients into axial-dominant versus sagittal-dominant phenotypes.

Biological insight. The observed coupling between axial rotation and compensatory lumbar lordosis supports the spiral column hypothesis of progressive scoliosis mechanics. Such coupling metrics could enrich risk-stratification models currently driven by Cobb angle alone.

Limitations. MRI acquisition is resource-intensive and not part of standard scoliosis work-ups in many centres. Supine positioning also attenuates deformity. Furthermore, the UK Biobank MRI's relatively coarse resolution (501×160×224 voxels) compared to clinical protocols (typically 1mm isotropic) may obscure subtle rotational features, particularly in mild curves where millimeter-scale asymmetries carry prognostic value. The homogeneous demographic of our training population (ages 40–69, healthy volunteer bias) limits applicability to pediatric idiopathic scoliosis or elderly osteoporotic populations where vertebral morphology differs substantially.

Literature Context and Clinical Impact. This chapter addresses the critical gap identified by [Zhang et al. 2019] and [Ma et al. 2020] regarding the neglect of axial and sagittal planes in scoliosis assessment. While previous 3D studies using EOS imaging were limited to $n = 100$ samples focusing on adolescent idiopathic scoliosis [Karam et al. 2022], we conducted the first population-scale 3D analysis on adult degenerative scoliosis. Our approach goes beyond the vertebrae-focused methods of [Chen et al. 2024] and [Li et al. 2024c] by analyzing the spine as an integrated 3D structure. This could change treatment planning: surgeons can now identify patients requiring derotation procedures versus simple coronal correction. The phenotypic clustering enables personalized treatment protocols. Axial-dominant cases may benefit from different bracing strategies than sagittal-dominant ones, moving beyond the one-size-fits-all approach criticized by [Smith et al. 2019].

Outlook. Ultra-fast, lower-field MRI (e.g. 0.55 T) could bridge cost gaps while preserving 3D fidelity. Federated training across vendors will be essential for generalisation.

8.3 Predicting 3D Spine Shape from a Single DXA View

Principal findings. To merge the accessibility of DXA with the anatomical fidelity of MRI, we introduce an image-based regression model that successfully maps a sin-

gle 2D DXA image to a 3D spine shape. A vision transformer learns this 2D to 3D correspondence, recreating the full spine shape geometry with sub-millimetre error.

Clinical implications. The model unlocks quasi-3D assessment using equipment already common in fracture-risk screening. The result is a practical bridge between modalities, offering clinical applicability in that DXA are cheap and more widely accessible unlike MRIs.

Limitations. The model inherits any DXA projection bias and assumes normal vertebral morphology. Most critically, training exclusively on UK Biobank's older adult cohort (mean age 55.8 years) introduces age-specific anatomical priors that increase reconstruction error when applied to younger populations, potentially compromising surgical planning accuracy in adolescent idiopathic scoliosis. The variable resolution of DXA inputs (0.6-1.2mm pixel spacing across different scanners) further compounds uncertainty in the 2D-to-3D mapping.

Literature Context and Clinical Impact. This represents a novel contribution not present in the transfer learning literature reviewed by [Morid et al. 2021] and [Constant et al. 2023]. While existing work focused on adapting pre-trained models within single modalities with SpineFM achieving 97.8% vertebrae identification [Simons et al. 2025] and MA-SAM demonstrating multi-atlas segmentation [Fan et al. 2024], we directly treated modality gap in training. Unlike the poor performance of SAM on spine MRI (IoU=0.1135) reported by [Mazurowski et al. 2023], our cross-modal synthesis maintains high fidelity. Diagnostically, this eliminates the binary choice between accessible but limited 2D imaging and comprehensive but expensive 3D imaging. Primary care physicians can now obtain 3D insights from routine DXA, enabling earlier referral decisions based on 3D deformity patterns rather than waiting for progression visible in 2D. This could shift intervention timing by 2 to 3 years earlier in the disease course, when conservative management is most effective.

Future work. A multi-view extension (antero-posterior + lateral) could further reduce depth ambiguity. Conditional variational formulations might output shape ensembles to express epistemic uncertainty when mapping from 2D to 3D.

8.4 Automated DXA Scoliosis Method (DSM)

Principal findings. When evaluated against 1,929 expert-annotated scans, DSM achieved an intraclass correlation coefficient of 0.89 for maximum angle and reduced reading time from ∼90s manually to 7s for automation.

Limitations and bias. The validation cohort skewed female (62%), reflecting osteoporosis screening demographics. Some failures trace back to patient positioning rather than algorithmic error. Resolution variations between clinical sites using different DXA manufacturers may introduce systematic measurement biases that could mask or exaggerate progression in longitudinal monitoring.

Literature Context and Clinical Impact. DSM directly addresses the inter-observer variability problem that has plagued manual measurement since Cobb's original 1948 method. While recent hybrid architectures like SpineHRformer [Zhao et al. 2023] focused on technical improvements in Cobb angle measurement, and CHASPPRAU-Net [Saeed et al. 2023] handled osteoporotic fractures, our work uniquely validates against large-scale clinical annotations. Unlike the technology-focused approaches of VertXNet's ensemble methods [Chen et al. 2024], we demonstrate real-world clinical concordance. Our automated DSM could transform diagnostic workflow a 93% time reduction in angle measurement. This could enable radiologists to focus on complex cases while te automated DSM handles routine screening. More critically, standardized measurements eliminate the 5-10° inter-observer variability that currently confounds progression monitoring. This standardization is essential for multi-centre clinical trials, where measurement inconsistency has been a major limitation identified by [Graaf et al. 2024] in the SPIDER challenge.

Next steps. We plan a prospective, multi-centre trial using our automated DSM as a triage tool, testing whether it reduces reported variations in human measurements.

8.5 Large-Scale, Generalisable 3D Spine Segmentation

Principal findings. Training on 51k MRI volumes yielded vertebra-level Dice of 0.95 and mean surface distance of 0.42mm.

Scalability. Using a hybrid Swin-Transformer/UNet backbone halved memory versus full-attention models, enabling full-field-of-view inputs at 1mm isotropic resolution on 24 GB GPUs. The pipeline now processes a UK Biobank MRI in 8s, fast enough for population-scale deployment.

Cross-dataset robustness. Despite strong performance within the UK Biobank, our segmentation model exhibits systematic degradation when applied to higher-resolution clinical datasets. The Dice score drops from 0.95 to 0.83 when tested on 1mm isotropic clinical MRI, primarily due to the model's training on UK Biobank's coarser $501 \times 160 \times 224$ resolution where fine anatomical boundaries are inherently blurred.

Literature Context and Clinical Impact. This chapter combines the efficiency goals of active learning [Budd et al. 2021], the label-efficiency of SSL methods like HD-Teacher [Zhu et al. 2023], and the architectural innovations of transformer-based models [Li et al. 2024c]. Our test-time adaptation module addresses a critical gap overlooked in prior work, namely the domain shift between training and test data. While SSHSNet achieved 96.12% Dice using cross pseudo supervision [Huang et al. 2023a] and HD-Teacher introduced dual 2D/3D frameworks, our unified approach achieves comparable performance while being deployable across heterogeneous scanners without retraining. Automated vertebra-level segmentation enables patient-specific finite element modeling for implant selection, potentially reducing the 3-5% implant failure rate. The 8s processing time allows real-time intraoperative segmentation for navigation systems, addressing the registration errors affecting spine surgeries.

Future directions. Incorporating self-supervised pre-training on unlabelled low-field MRI could cut annotation further. Shape-aware diffusion models may regularise segmentations under extreme pathology while preserving fine anomalies.

8.6 Impact of Resolution and Population Heterogeneity on Model Performance

Resolution variability across imaging modalities. The substantial resolution heterogeneity across our datasets from DXA scans 832×320 to the standardized $240 \times 240 \times 155$ voxels of BRATS and $501 \times 160 \times 224$ of UK Biobank MRI presents fundamental chal-

lenges for model generalization. DXA's planar projection fundamentally differs from the volumetric nature of MRI, with effective resolution varying across the imaging field due to fan-beam geometry. The UK Biobank MRI's relatively coarse resolution ($501 \times 160 \times 224$) compared to specialized datasets like BRATS ($240 \times 240 \times 155$ at 1mm isotropic) impacts fine anatomical detail capture, particularly for small vertebral features critical in early deformity detection. Our models must therefore learn resolution-invariant representations, which we is achieved through multi-scale feature extraction in the Swin-Transformer backbone (Chapter 7) and progressive upsampling in the 2D-to-3D synthesis network (Chapter 5).

Population demographic shifts. The demographic composition varies dramatically across our datasets, introducing systematic biases that affect model performance. The UK Biobank cohorts (both DXA and MRI) represent a predominantly white British population aged 40 to 69 with healthy volunteer bias, contrasting sharply with the BRATS dataset's international brain tumor patients spanning all age groups. ALSPAC's longitudinal pediatric cohort introduces age-related anatomical variation absent in adult-focused datasets. This population heterogeneity manifests in several ways: (i) *Anthropometric differences:* UK Biobank's European-centric cohort may exhibit different vertebral dimensions and curvature patterns compared to a more diverse population, (ii) *Disease spectrum bias:* while UK Biobank captures primarily degenerative scoliosis in older adults, clinical datasets like BRATS and AMOS include pathological cases with severe anatomical distortion that challenge segmentation assumptions, (iii) *Socioeconomic confounding* UK Biobank's volunteer bias toward higher socioeconomic status correlates with better baseline spine health, potentially inflating performance metrics when models trained on this data encounter real-world clinical populations.

Mitigation strategies employed. To address these challenges, we implemented several domain adaptation techniques: (i) *Resolution harmonization:* all volumetric data underwent standardized preprocessing including resampling to common spacing and intensity normalization, though this risks losing high-frequency details in originally high-resolution scans, (ii) *Test-time adaptation* (Chapter 7): our self-recalibration module adjusts batch normalization statistics during inference, partially compensating for population shifts with 6% Dice improvement on out-of-distribution data.

Limitations of current approaches. Despite these mitigations, fundamental limitations persist. Resolution harmonization through resampling cannot recover information absent in low-resolution acquisitions. Sub-millimeter cortical features visible in BRATS remain inaccessible in UK Biobank MRI. Population-specific anatomical variations, particularly in understudied groups (e.g., individuals of African or South Asian ancestry, pediatric populations with growth disorders), remain poorly captured due to dataset composition. The healthy volunteer bias in UK Biobank means our "normal" spine models may systematically underperform on clinical populations with comorbidities affecting bone quality (osteoporosis, metastases, inflammatory arthropathies).

Clinical implications of heterogeneity. These resolution and population effects have direct clinical consequences. Models trained on UK Biobank's relatively healthy population may miss subtle early degenerative changes when deployed in osteoporosis clinics. The shape reconstruction error could potentially increase in younger populations and misguide surgical planning in adolescent idiopathic scoliosis, where millimeter precision affects fusion levels. Resolution limitations mean that while our methods excel at detecting moderate-to-severe deformity (maximum spine angle $> 20°$), they may miss mild curves ($10 - 20°$) that benefit most from early intervention. Geographic deployment must therefore be carefully validated. A model performing well in UK teaching hospitals may require recalibration for rural clinics serving different ethnic compositions or using older, lower-resolution imaging equipment.

Future directions for robust generalization. Addressing these challenges requires both technical and infrastructural advances: (i) *Federated learning initiatives* could train models across diverse populations without centralizing sensitive data, though heterogeneous imaging protocols complicate implementation; (ii) *Super-resolution networks* specifically designed for medical imaging could bridge resolution gaps, though hallucinated details risk clinical misinterpretation; (iii) *Domain-invariant learning* using adversarial training or optimal transport could learn features robust to population shifts, at the cost of potentially reduced performance on any single domain; (iv) *Prospective multi-center validation* with predefined demographic strata would quantify real-world generalization gaps and guide targeted data collection; (v) *Physics-informed architectures* that explicitly model imaging system characteristics (DXA fan-beam geometry, MRI coil sensitivity profiles) could separate true anatomical variation from acquisition artifacts.

8.7 Synthesis and Future Directions

Connecting the pieces. Chapters 2 to 6 form a pipeline that (i) *ingests* ubiquitous DXA, (ii) *augments* it to MRI-grade 3D insight, and (iii) *scales* to national biobank. Together they answer why earlier work falls short in terms of: limited labels, 2D only measurements, and modality specific models. Each chapter incrementally relaxes these constraints.

Population-scale epidemiology. Deploying DSM over all 48,384 UK Biobank DXA scans will, for the first time, quantify adult scoliosis prevalence with narrow confidence intervals. Coupled with genetic data, this enables genome-wide association studies on curvature phenotypes.

Personalised intervention. 3D shape-aware prognostic models could predict curve progression risk on a per-patient basis, informing earlier bracing or surgical timing decisions. Integrating our shape descriptors with finite-element simulations opens avenues for patient-specific biomechanical planning.

Ethical and societal considerations. Automated triage risks overburdening secondary care if false positives spike referrals. Continuous audit dashboards reporting sensitivity, specificity, and demographic parity must accompany widescale roll-out. Secure, federated analytics can respect privacy while enabling cross-centre learning.

Long-term vision. We envisage a learning health-system loop where each new scan refines population statistics, uncertainty estimates prompt targeted labelling, and clinicians receive adaptive, interpretable decision support for realising truly continuous, data-driven scoliosis management. Crucially, this system must actively monitor and correct for demographic and technical biases, automatically flagging when model confidence drops due to population or resolution mismatches, and preferentially requesting annotations from underrepresented groups to continuously improve equity in model performance across all patient populations and imaging contexts.

References

Aaro, Stig, Dahlborn, Mats, and Svensson, Leif (1978). "Estimation of Vertebral Rotation in Structural Scoliosis by Computer Tomography". In: *Acta Radiologica*. DOI: 10.1177/028418517801900614.

Akinci D'Antonoli, Tugba, Berger, Lucas K., Indrakanti, Ashraya K., Vishwanathan, Nathan, Weiss, Jakob, Jung, Matthias, Berkarda, Zeynep, Rau, Alexander, Reisert, Marco, Küstner, Thomas, Walter, Alexandra, Merkle, Elmar M., Boll, Daniel T., Breit, Hanns-Christian, Nicoli, Andrew Phillip, Segeroth, Martin, Cyriac, Joshy, Yang, Shan, and Wasserthal, Jakob (2025). "TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI". In: *Radiology*. PMID: 39964271. DOI: 10.1148/radiol.241613.

Aubert, Benjamin, Vazquez, Carlos, Cresson, Thierry, Parent, Stefan, and Guise, Jacques A de (2019). "Toward automated 3D spine reconstruction from biplanar radiographs using CNN for statistical spine model fitting". In: *Transaction on Medical Imaging (TMI)*. DOI: 10.1109/TMI.2019.2914400.

Baid, Ujjwal, Ghodasara, Satyam, Bilello, Michel, Mohan, Suyash, Calabrese, Evan, Colak, Errol, Farahani, Keyvan, Kalpathy-Cramer, Jayashree, Kitamura, Felipe Campos, Pati, Sarthak, Prevedello, Luciano M., Rudie, Jeffrey D., Sako, Chiharu, Shinohara, Russell T., Bergquist, Timothy, Chai, Rong, Eddy, James A., Elliott, Julia, Reade, Walter Caswell, Schaffter, Thomas, Yu, Thomas, Zheng, Jiaxin, Annotators, BraTS, Davatzikos, Christos, Mongan, John T, Hess, Christopher Paul, Cha, Soonmee, Villanueva-Meyer, Javier E., Freymann, John B., Kirby, Justin S., Wiestler, Benedikt, Crivellaro, Priscila Sacilotto, R.Colen, Rivka, Kotrotsou, Aikaterini, Marcus, Daniel, Milchenko, Mikhail, Nazeri, Arash, Fathallah-Shaykh, Hassan M., Wiest, Roland, Jakab, András, Weber, Marc-André, Mahajan, Abhishek, Menze, Bjoern H, Flanders, Adam E., and Bakas, Spyridon (2021). "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification". In: *ArXiv*.

Bakas, Spyridon, Akbari, Hamed, Sotiras, Aristeidis, Bilello, Michel, Rozycki, Martin, Kirby, Justin S., Freymann, John B., Farahani, Keyvan, and Davatzikos, Christos (2017). "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". English (US). In: *Scientific Data*. DOI: 10.1038/sdata.2017.117.

Baur, Christoph, Albarqouni, Shadi, and Navab, Nassir (2017). "Semi-Supervised Deep Learning for Fully Convolutional Networks". In: *Deep Learning and Data Labeling for Medical Applications*. Lecture Notes in Computer Science. Springer. DOI: 10.1007/978-3-319-66179-7_36.

Benameur, Said, Mignotte, Max, Parent, Stefan, Labelle, Hubert, Skalli, Wafa, and Guise, Jacques de (2003). "3D/2D registration and segmentation of scoliotic vertebrae using statistical models". In: *Computerized Medical Imaging and Graphics*. DOI: 10.1016/s0895-6111(03)00019-3.

Boehringer, Andrew S., Sanaat, Amirhossein, Arabi, Hossein, and Zaidi, Habib (2023). "An active learning approach to train a deep learning algorithm for tumor segmentation from brain MR images". In: *Insights into Imaging*. DOI: 10.1186/s13244-023-01487-6.

Bourigault, Emmanuelle, Hamdi, Abdullah, and Jamaludin, Amir (2024a). *X-Diffusion: Generating Detailed 3D MRI Volumes From a Single Image Using Cross-Sectional Diffusion Models*.

Bourigault, Emmanuelle, Jamaludin, Amir, Clark, Emma, Fairbank, Jeremy, Kadir, Timor, and Zisserman, Andrew (2023). "3D Shape Analysis of Scoliosis". In: *MICCAI Workshop on Shape in Medical Imaging*. DOI: 10.1007/978-3-031-46914-5_22.

Bourigault, Emmanuelle, Jamaludin, Amir, and Hamdi, Abdullah (2025). *UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation*.

Bourigault, Emmanuelle, Jamaludin, Amir, Kadir, Timor, and Zisserman, Andrew (2022). "Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach". In: *Medical Imaging with Deep Learning*.

Bourigault, Emmanuelle, Jamaludin, Amir, and Zisserman, Andrew (2024b). "3D Spine Shape Estimation from Single 2D DXA". In: *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. DOI: 10.1007/978-3-031-72086-4_1.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., and Davey Smith, G. (2013). "Cohort profile: the 'children of the 90s'—the index offspring of the Avon longitudinal study of parents and children". In: *International Journal of Epidemiology*. DOI: 10.1093/ije/dys064.

Budd, Samuel, Robinson, Emma C., and Kainz, Bernhard (2021). "A survey on active learning and human-in-the-loop deep learning for medical image analysis". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2021.102062.

Carman, D. L., Browne, R. H., and Birch, J. G. (1990). "Measurement of scoliosis and kyphosis radiographs. Intraobserver and interobserver variation". In: *Journal of Bone and Joint Surgery American Volume*. DOI: 10.2106/00004623-199072030-00003.

Carman, DL, Browne, RH, and Birch, JG (1989). "Interobserver and intraobserver reliability of the Cobb angle". In: *Journal of Pediatric Orthopedics*.

Carreau, Joseph H, Bastrom, Tracey P., Petcharaporn, Maty, Schulte, Caitlin, Marks, Michelle M, Illés, Tamás S., Somoskeöy, Szabolcs, and Newton, Peter O. (2014). "Computer-Generated, Three-Dimensional Spine Model From Biplanar Radiographs: A Validity Study in Idiopathic Scoliosis Curves Greater Than 50 Degrees". In: *Spine Deformity*. DOI: 10.1016/j.jspd.2013.10.003.

Castelein, René M. (2012). "Pre-existent Rotation of the Normal Spine at Different Ages and its Consequences for the Scoliotic Mechanism". In: *Research into Spinal Deformities 8*. Studies in Health Technology and Informatics. IOS Press. DOI: 10.3233/978-1-61499-067-3-20.

Chen, Baixu, Jiang, Junguang, Wang, Ximei, Wan, Pengfei, Wang, Jianmin, and Long, Mingsheng (2022). "Debiased self-training for semi-supervised learning". In: *Advances in Neural Information Processing Systems*.

Chen, Jianpeng, Lu, Yongyi, Yu, Qihang, Luo, Yan, Zhou, Yuyin, Kalantari, Nader Dagley, Fan, Zhe, Lu, Wenqi, Rhee, Daehyun, El Fakhri, Georges, Fischl, Bruce, Frangi, Alejandro F., Golland, Polina, Guo, Hengtao, Sun, Liang, and Zhang, Zhuangzhi (2021). "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*. Lecture Notes in Computer Science.

Chen, Runnan, Liu, Youquan, Kong, Lingdong, Zhu, Xinge, Ma, Yuexin, Li, Yikang, Hou, Yuenan, Qiao, Yu, and Wang, Wenping (2023). "CLIP2Scene: Towards Label-Efficient 3D Scene Understanding by CLIP". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR52729.2023.00678.

Chen, Y, Mo, Y, Readie, A, et al. (2024). "VertXNet: an ensemble method for vertebral body segmentation and identification from cervical and lumbar spinal X-rays". In: *Scientific Reports*. DOI: 10.1038/s41598-023-49923-3.

Chen, Yuhua, Ruan, Dan, Xiao, Jiayu, Wang, Lixia, Sun, Bin, Saouaf, Rola, Yang, Wensha, Li, Debiao, and Fan, Zhaoyang (2020). "Fully Automated Multi-Organ Segmentation in

Abdominal Magnetic Resonance Imaging with Deep Neural Networks". In: *Medical physics*. DOI: 10.1002/MP.14429.

Cheng, Junlong, Fu, Bin, Ye, Jin, Wang, Guoan, Li, Tianbin, Wang, Haoyu, Li, Ruoyu, Yao, He, Chen, Junren, Li, Jingwen, Su, Yanzhou, Zhu, Min, and He, Junjun (2025). "Interactive Medical Image Segmentation: A Benchmark Dataset and Baseline". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, Peng, Yang, Yongqiang, Yu, Haifeng, and He, Yupeng (2021). "Automatic vertebrae localization and segmentation in CT with a two-stage Dense-U-Net". In: *Scientific Reports*. DOI: 10.1038/s41598-021-01296-1.

Cheplygina, Veronika, deBruijne, Marleen, and Pluim, Josien P.W. (2019). "Not-So-Supervised: A Survey of Semi-Supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2019.03.009.

Cheung, Jason Pui Yin, Cheung, Kenneth Man Chee, Samartzis, Dino, and Luk, Keith Dip Kei (2022). "Genetic determinants of adolescent idiopathic scoliosis progression". In: *Nature Reviews Rheumatology*.

Choy, Christopher, Gwak, JunYoung, and Savarese, Silvio (2019). "4d spatio-temporal convnets: Minkowski convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Clark, Eileen M, Taylor, Hugh J, Harding, Ian, Hutchinson, John, Nelson, Ian, Deanfield, John E, and Tobias, Jon H (2014). "Differential vertebral growth patterns in adolescent idiopathic scoliosis". In: *Journal of Bone and Mineral Research*.

Clogenson, Marine, Duff, John M, Luethi, Marcel, Levivier, Marc, Meuli, Reto, Baur, Charles, and Henein, Simon (2015). "A statistical shape model of the human second cervical vertebra". In: *IJCARS*. DOI: 10.1007/s11548-014-1121-x.

Cobb, John Robert (1948). "Outline for the study of scoliosis". In: *American Academy of Orthopedic Surgeons Instructional Course Lectures*.

Constant, Caroline, Aubin, Carl-Eric, Kremers, Hilal Maradit, Garcia, Diana V. Vera, Wyles, Cody C., Rouzrokh, Pouria, and Larson, Annalise Noelle (Sept. 2023). "The use of deep learning in medical imaging to improve spine care: A scoping review of current literature and clinical applications". In: *North American Spine Society Journal*. DOI: 10.1016/j.xnsj.2023.100236.

Cootes, Timothy F., Taylor, Christopher J., Cooper, David H., and Graham, Jim (Jan. 1995). "Active Shape Models—Their Training and Application". In: *Computer Vision and Image Understanding*. DOI: 10.1006/cviu.1995.1004.

Courvoisier, Aurélien, Vialle, Raphaël, and Skalli, Wafa (2014). “EOS 3D Imaging: assessing the impact of brace treatment in adolescent idiopathic scoliosis”. In: *Expert Review of Medical Devices*.

Cristante, Alexandre Fogaça, Silva, Ricardo Teixeira e, Costa, Guilherme Henrique Ricardo da, and Marcon, Raphael Martus (2021). “Adult degenerative scoliosis”. In: *Revista Brasileira de Ortopedia*. DOI: 10.1055/s-0040-1709736.

D’Andrea, Luca et al. (2021). “Automated segmentation of spinal structures in MRI: Techniques and challenges”. In: *IEEE Transactions on Medical Imaging*.

Daniel, Rui, Barbosa, Daniel, and Cardoso, Jaime S. (2025). “Continual Deep Active Learning for Medical Imaging: Replay-Based Architecture for Context Adaptation”. In: *arXiv preprint arXiv:2501.08245*. URL: https://arxiv.org/abs/2501.08245.

Yi-de, Ma, Qing, Liu, and Zhi-bai, Qian (2004). “Automated image segmentation using improved PCNN model based on cross-entropy”. In: *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*

Ding, Runyu, Yang, Jihan, Xue, Chuhui, Zhang, Wenqing, Bai, Song, and Qi, Xiaojuan (2023). “PLA: Language-Driven Open-Vocabulary 3D Scene Understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong, Haoyu, Konz, Nicholas, Gu, Hanxue, and Mazurowski, Maciej A. (June 2024). “Medical Image Segmentation with InTEnt: Integrated Entropy Weighting for Single Image Test-Time Adaptation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Donzelli, Sabrina, Poma, Salvatore, Balzarini, Luca, Borboni, Alberto, Respizzi, Stefano, Villafañe, Jorge, Zaina, Fabio, and Negrini, Stefano (Oct. 2015). “State of the art of current 3-D scoliosis classifications: a systematic review from a clinical perspective”. In: *Journal of NeuroEngineering and Rehabilitation (JNER)*. DOI: 10.1186/s12984-015-0083-8.

Doran, Simon J., Hipwell, John H., Denholm, Rachel, Eiben, Björn, Busana, Marta, Hawkes, David J., Leach, Martin O., and Santos Silva, Isabel dos (2017). “Breast MRI segmentation for density estimation: Do different methods give the same results and how much do differences matter?” In: *Medical Physics*. Epub 2017 Jul 25. DOI: 10.1002/mp.12320.

Dubousset, Jean, Ilharreborde, Brice, and Huec, Jean-Charles Le (2014). “Use of EOS imaging for the assessment of scoliosis deformities: application to postoperative 3D quantitative analysis of the trunk”. In: *European Spine Journal*.

Ezhov, N., Neitzel, F., and Petrovic, S. (2018). “Spline approximation, part 1: Basic methodology”. In: *Journal of Applied Geodesy*.

Fan, Dongping et al. (2024). “MA-SAM: A Multi-atlas Guided SAM Using Pseudo Mask Prompts without Manual Annotation for Spine Image Segmentation”. In: *arXiv preprint*.

Fang, Xi and Yan, Pingkun (2020). “Multi-Organ Segmentation Over Partially Labeled Datasets With Multi-Scale Feature Abstraction”. In: *IEEE Transactions on Medical Imaging*. DOI: 10.1109/TMI.2020.3001036.

Fiorentino, Maria Chiara, Villani, Francesca Pia, Benito Herce, Rocío, González Ballester, Miguel Angel, Mancini, Adriano, and López-Linares Román, Karen (2024). “An intensity-based self-supervised domain adaptation method for intervertebral disc segmentation in magnetic resonance imaging”. In: *International Journal of Computer Assisted Radiology and Surgery*.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S. M., and Lawlor, D. A. (2013). “Cohort profile: the Avon longitudinal study of parents and children: ALSPAC mothers cohort”. In: *International Journal of Epidemiology*. DOI: 10.1093/ije/dys066.

Gajny, Laurent, Ebrahimi, Shahin, Vergari, Claudio, Angelini, Elsa, and Skalli, Wafa (2019). “Quasi-automatic 3D reconstruction of the full spine from low-dose biplanar X-rays based on statistical inferences and image analysis”. In: *European Spine Journal*. Epub 2018 Oct 31. DOI: 10.1007/s00586-018-5807-6.

Galbusera, Fabio, Bassani, Tito, Panico, Matteo, Sconfienza, Luca Maria, and Cina, Andrea (2022). “A fresh look at spinal alignment and deformities: Automated analysis of a large database of 9832 biplanar radiographs”. In: *Frontiers in Bioengineering and Biotechnology*.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron (2016). *Deep Learning*. MIT Press.

Graaf, Jasper W. van der, Hooff, Miranda L. van, Buckens, Constantinus F.M., Rutten, Matthieu, Susante, Job L.C. van, Kroeze, Robert Jan, Kleuver, Marinus de, Ginneken, Bram van, and Lessmann, Nikolas (2024). “Lumbar spine segmentation in MR images: A dataset and a public benchmark”. In: *Scientific Data*.

Graf, Robert, Platzek, Paul-Sören, Riedel, Evamaria Olga, Ramschütz, Constanze, Starck, Sophie, Möller, Hendrik Kristian, Atad, Matan, Völzke, Henry, Bülow, Robin, Schmidt, Carsten Oliver, et al. (2024). “TotalVibeSegmentator: Full Body MRI Segmentation for the NAKO and UK Biobank”. In: *arXiv preprint arXiv:2406.00125*.

Gstoettner, M., Sekyra, K., Walochnik, N., Winter, P., Wachter, R., and Bach, C. M. (2007). “Inter- and intraobserver reliability assessment of the Cobb angle: manual versus digital measurement tools”. In: *European Spine Journal*. DOI: 10.1007/s00586-007-0401-3.

Ha, Huy and Song, Shuran (2022). “Semantic Abstraction: Open-World 3D Scene Understanding from 2D Vision-Language Models”. In: *Proceedings of the Conference on Robot Learning (CoRL)*.

Hamdi, Abdullah, Ghanem, Bernard, and Nießner, Matthias (2023a). “Sparf: Large-scale learning of 3d sparse radiance fields from few input images”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Hamdi, Abdullah, Giancola, Silvio, and Ghanem, Bernard (Oct. 2021). “MVTN: Multi-View Transformation Network for 3D Shape Recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hamdi, Abdullah, Giancola, Silvio, and Ghanem, Bernard (2023b). “Voint Cloud: Multi-View Point Cloud Representation for 3D Understanding”. In: *The Eleventh International Conference on Learning Representations*.

Hatamizadeh, Ali, Nath, Vishwesh, Tang, Yucheng, Yang, Dong, Roth, Holger R., and Xu, Daguang (2022). “Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021), MICCAI Workshop*. Lecture Notes in Computer Science. Springer. DOI: 10.1007/978-3-031-08999-2_22.

Hatamizadeh, Ali, Yang, Dong, Roth, Holger R., and Xu, Daguang (2021). “UNETR: Transformers for 3D Medical Image Segmentation”. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Held, Jan, Cioppa, Anthony, Giancola, Silvio, Hamdi, Abdullah, Ghanem, Bernard, and Van Droogenbroeck, Marc (2023). “VARS: Video Assistant Referee System for Automated Soccer Decision Making from Multiple Views”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ho, E K, Upadhyay, Shanti S., Chan, F. L., Hsu, Louis C. S., and Leong, J. C. Y. (1993). “New methods of measuring vertebral rotation from computed tomographic scans. An intraobserver and interobserver study on girls with scoliosis.” In: *Spine*.

Hu, Zongshan, Vergari, Claudio, Gajny, Laurent, Liu, Zhen, Lam, Tsz-Ping, Zhu, Zezhang, Qiu, Yong, Man, Gene C. W., Yeung, Kwong-Hang, Chu, Winnie C. W., Cheng, Jack C. Y., and Skalli, Wafa (2021). “Comparison of 3D and 2D characterization of spinal geometry from biplanar X-rays: a large cohort study”. In: *Quantitative Imaging in Medicine and Surgery*. URL: https://qims.amegroups.org/article/view/66766.

Huang, Meiyan, Zhou, Shuoling, Chen, Xiang, Lai, Haoran, and Feng, Qianjin (2023a). “Semi-supervised hybrid spine network for segmentation of spine MR images”. In: *Computerized Medical Imaging and Graphics*. DOI: 10.1016/j.compmedimag.2023.102245.

Huang, Tianyu, Dong, Bowen, Yang, Yunhan, Huang, Xiaoshui, Lau, Rynson W. H., Ouyang, Wanli, and Zuo, Wangmeng (2023b). "CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ilharreborde, B., Steffen, J. S., Nectoux, E., Vital, J. M., Mazda, K., Skalli, W., and Obeid, I. (Sept. 2011). "Angle measurement reproducibility using EOS three-dimensional reconstructions in adolescent idiopathic scoliosis treated by posterior instrumentation". In: *Spine (Phila Pa 1976)*. DOI: DOI:10.1097/BRS.0b013e3182293548.

Illés, Tamás S., Lavaste, F., and Dubousset, Jean (2019). "The third dimension of scoliosis: The forgotten axial plane." In: *OTSR*.

Illés, Tamás S., Tunyogi-Csapó, Miklós, and Somoskeöy, Szabolcs (2010). "Breakthrough in three-dimensional scoliosis diagnosis: significance of horizontal plane view and vertebra vectors". In: *European Spine Journal*. DOI: DOI:10.1007/s00586-010-1566-8.

Irvin, Jeremy, Rajpurkar, Pranav, Ko, Michael, Yu, Yifan, Ciurea-Ilcus, Silviana, Chute, Chris, Marklund, Henrik, Haghgoo, Behzad, Ball, Robyn, Shpanskaya, Kseniya, Seekins, Jane, Mong, Dima A., Halabi, Safwan S., Sandberg, Jonas K., Jones, Russell, Larson, David B., Langlotz, Curtis P., Patel, Bhavik N., Lungren, Matthew P., and Ng, Andrew Y. (2019). "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. DOI: 10.1609/aaai.v33i01.3301590.

Isensee, Fabian, Jaeger, Paul F, Kohl, Simon A A, Petersen, Jens, and Maier-Hein, Klaus H (Feb. 2021). "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods*. DOI: 10.1038/s41592-020-01008-z.

Jamaludin, A., Kadir, T., Clark, E., and Zisserman, A. (2019a). "Predicting Spine Geometry and Scoliosis from DXA Scans". In: *MICCAI Workshop: MSKI*. DOI: 10.1007/978-3-030-13736-6_2.

Jamaludin, Amir, Kadir, Timor, Clark, Emma, and Zisserman, Andrew (2019b). "Predicting Spine Geometry and Scoliosis from DXA Scans". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI) Workshop: Computational Methods and Clinical Applications in Musculoskeletal Imaging*.

Jamaludin, Amir, Kadir, Timor, Clark, Emma M., and Zisserman, Andrew (2018). "Predicting Scoliosis in DXA Scans Using Intermediate Representations". In: *CSI Medical Image Computing and Computer Assisted Intervention*. DOI: 10.1007/978-3-030-13736-6_2.

Jamaludin, Amir, Kadir, Timor, and Zisserman, Andrew (2017). "Self-supervised Learning for Spinal MRIs". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for*

Clinical Decision Support (DLMIA ML-CDS 2017). Lecture Notes in Computer Science. Springer. DOI: 10.1007/978-3-319-67558-9_34.

Jamaludin, Amir, Kadir, Timor, and Zisserman, Andrew (2020). "Automated Cobb angle measurement using deep learning". In: *Medical Image Analysis*.

Jamaludin, Amir, Lootus, Meelis, Kadir, Timor, and Zisserman, Andrew (2016). "Automatic Intervertebral Discs Localization and Segmentation: A Vertebral Approach". In: *CSI proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*. DOI: 10.1007/978-3-319-41827-8_9.

Jha, Debesh, Smedsrud, Pia Helen, Riegler, M., Johansen, Dag, Lange, Thomas de, Halvorsen, P., and Johansen, Håvard Dagenborg (2019). "ResUNet++: An Advanced Architecture for Medical Image Segmentation". In: *2019 IEEE International Symposium on Multimedia (ISM)*. DOI: 10.1109/ISM46123.2019.00049.

Ji, Yuanfeng, Bai, Haotian, Ge, Chongjian, Yang, Jie, Zhu, Ye, Zhang, Ruimao, Li, Zhen, Zhang, Lingyan, Ma, Wanling, Wan, Xiang, and Luo, Ping (2022). "AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation". In: *Advances in Neural Information Processing Systems*. Ed. by Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. Curran Associates, Inc.

Johnson, Alistair E. W., Pollard, Tom J., Berkowitz, Seth J., Greenbaum, Nathaniel R., Lungren, Matthew P., Deng, Chih-ying, Mark, Roger G., and Horng, Steven (2019). "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data*. DOI: 10.1038/s41597-019-0322-0.

Jones, Michael et al. (2018). "Deep learning approaches for spinal MRI segmentation: A review". In: *Journal of Digital Imaging*.

Karam, Nadine et al. (2022). "Global morphological insights into scoliosis: A UKBiobank study". In: *Medical Image Analysis*.

Karani, Neerav, Erdil, Ertunc, Chaitanya, Krishna, and Konukoglu, Ender (2021). "Test-time adaptable neural networks for robust medical image segmentation". In: *Medical Image Analysis*. DOI: https://doi.org/10.1016/j.media.2020.101907.

Karimi, Davood and Salcudean, Septimiu E (2019). "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks". In: *IEEE Transactions on medical imaging*.

Kerr, Justin, Kim, Chung Min, Goldberg, Ken, Kanazawa, Angjoo, and Tancik, Matthew (2023). "LERF: Language Embedded Radiance Fields". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Khalil, Yasmina Al, Becherucci, Edoardo A, Kirschke, Jan Stefan, Karampinos, Dimitrios C., Breeuwer, Marcel M., Baum, Thomas, and Sollmann, Nico (2022). "Multi-scanner and

multi-modal lumbar vertebral body and intervertebral disc segmentation database". In: *Scientific Data*.

Kilshaw, M., Baker, R. P., Gardner, R., Charosky, S., and Harding, I. (2019). "Abnormalities of the lumbar spine in the coronal plane on plain abdominal radiographs". In: *European Spine Journal*. DOI: 10.1007/s00586-010-1610-8.

Kim, Hee E., Cosa-Linan, Alejandro, Santhanam, Nandhini, Jannesari, Mahboubeh, Maros, Mate E., and Ganslandt, Thomas (2022). "Transfer learning for medical image classification: a literature review". In: *BMC Medical Imaging*. DOI: 10.1186/s12880-022-00793-7.

Kingma, Diederik P. and Ba, Jimmy (2015). "Adam: A Method for Stochastic Optimization". In: *ICLR*.

Kirillov, Alexander, Mintun, Eric, Ravi, Nikhila, Mao, Hanzi, Rolland, Chloe, Gustafson, Laura, Xiao, Tete, Whitehead, Spencer, Berg, Alexander C., Lo, Wan-Yen, et al. (2023). "Segment Anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Kobayashi, Sosuke, Matsumoto, Eiichi, and Sitzmann, Vincent (2022). "Decomposing Nerf for Editing via Feature Field Distillation". In: *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Komeili, Amin, Parent, Eric C, El-Rich, Marwan, and Adeeb, Samer (2019). "3D analysis of scoliosis progression using digital templating". In: *European Spine Journal*.

Konieczny, Markus Rafael, Senyurt, Hüsseyin, and Krauspe, Rüdiger (2013). "Epidemiology of adolescent idiopathic scoliosis". In: *Journal of Children's Orthopaedics*.

Kouwenhoven, Jan W., Van Ommeren, Peter M., Pruijs, Hans E. J., and Castelein, René M. (2006). "Spinal Decompensation in Neuromuscular Disease". In: *Spine*. DOI: 10.1097/01.brs.0000208131.42824.c3.

Kuş, Zeki and Aydin, Musa (2024). "MedSegBench: A comprehensive benchmark for medical image segmentation in diverse data modalities". In: *Scientific Data*.

Langensiepen, S, O, Semler, R, Sobottke, O, Fricke, J, Franklin, E, Schönau, and P, Eysel (2013). "Measuring procedures to determine the Cobb angle in idiopathic scoliosis: a systematic review." In: *Eur Spine J.*

Larios, Francisco, Mayer, Alexandria, Jimenez, Juan Nicolas Barajas, Morgan, Steven J., Khalil, Safa, Buckland, Aaron J., Protopsaltis, Themistocles S., Lafage, Virginie, Kim, Han Jo, and Schwab, Frank J. (2024). "Adolescent idiopathic scoliosis in adulthood". In: *EFORT Open Reviews*.

Lee, Ho Hin, Bao, Shunxing, Huo, Yuankai, and Landman, Bennett (2023). "3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image

Segmentation". In: *International Conference on Learning Representations*. DOI: 10.48550/arXiv.2209.15076.

Li, Keyu, Gu, Hanxue, Colglazier, Roy, Lark, Robert, Hubbard, Elizabeth, French, Robert, Smith, Denise, Zhang, Jikai, McCrum, Erin, Catanzano, Anthony, Cao, Joseph, Waldman, Leah, Mazurowski, Maciej A., and Alman, Benjamin (2024a). "Deep Learning Automates Cobb Angle Measurement Compared with Multi-Expert Observers". In: *arXiv preprint*.

Li, Wenxuan, Qu, Chongyu, Chen, Xiaoxi, Bassi, Pedro R. A. S., Shi, Yijia, Lai, Yuxiang, Yu, Qian, Xue, Huimin, Chen, Yixiong, Lin, Xiaorui, Tang, Yutong, Cao, Yining, Han, Haoqi, Zhang, Zheyuan, Liu, Jiawei, Zhang, Tiezheng, Ma, Yujiu, Wang, Jincheng, Zhang, Guang, Yuille, Alan, and Zhou, Zongwei (Oct. 2024b). "AbdomenAtlas: A Large-Scale, Detailed-Annotated, & Multi-Center Dataset for Efficient Transfer Learning and Open Algorithmic Benchmarking". In: *Medical Image Analysis*.

Li, X, Hong, Y, Xu, Y, and Hu, M (2024c). "VerFormer: Vertebrae-Aware Transformer for Automatic Spine Segmentation from CT Images". In: *Diagnostics*. DOI: 10.3390/diagnostics14171859.

Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, Laak, Jeroen AWM van der, Ginneken, Bram van, and Sánchez, Clara I (2017). "A Survey on Deep Learning in Medical Image Analysis". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2017.07.005.

Liu, Kaiping, Mokhtari, Mehdi, Li, Bo, Nofallah, Shima, Wu, Xiaomin, Ghorbani, Ali, and Shen, Luyao (2022). "Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning". In: *Medical Image Analysis*.

Lonstein, John E and Carlson, John M (1984). "The prediction of curve progression in untreated idiopathic scoliosis". In: *Journal of Bone and Joint Surgery*.

López Picazo, Mirella, Magallón Baro, Alba, Del Río Barquero, Luis M., Di Gregorio, Silvana, Martelli, Yves, Romera, Jordi, Steghöfer, Martin, González Ballester, Miguel A., and Humbert, Ludovic (2018). "3-D Subject-Specific Shape and Density Estimation of the Lumbar Spine From a Single Anteroposterior DXA Image Including Assessment of Cortical and Trabecular Bone". In: *Transaction on Medical Imaging (TMI)*. DOI: 10.1109/TMI.2018.2845909.

Lu, Yuheng, Xu, Chenfeng, Wei, Xiaobao, Xie, Xiaodong, Tomizuka, Masayoshi, Keutzer, Kurt, and Zhang, Shanghang (2023). "Open-Vocabulary Point-Cloud Object Detection without 3D Annotation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, Jun, He, Yuting, Li, Feifei, Han, Lin, You, Chenyu, and Wang, Bo (2024). "Segment anything in medical images". In: *Nature Communications*. DOI: 10.1038/s41467-024-44824-z.

Ma, Qichao, Wang, Lin, Zhao, Lihua, Wang, Yicheng, Chen, Mengjie, Wang, Sun, Lv, Zhibao, and Luo, Yi (2020). "Coronal Balance vs. Sagittal Profile in Adolescent Idiopathic Scoliosis, Are They Correlated?" In: *Frontiers in Pediatrics*. DOI: 10.3389/fped.2019.00523.

Mai, Jinjie, Hamdi, Abdullah, Giancola, Silvio, Zhao, Chen, and Ghanem, Bernard (2023). "Egoloc: Revisiting 3d object localization from egocentric videos with visual queries". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Mai, Jinjie, Zhu, Wenxuan, Rojas, Sara, Zarzar, Jesus, Hamdi, Abdullah, Qian, Guocheng, Li, Bing, Giancola, Silvio, and Ghanem, Bernard (2024). "TrackNeRF: Bundle Adjusting NeRF from Sparse and Noisy Views via Feature Tracks". In: *Computer Vision – ECCV 2024*. Lecture Notes in Computer Science. Springer. DOI: 10.1007/978-3-031-73254-6_27.

Marty-Poumarat, Catherine, Scattin, Laurent, Marpeau, Marc, Garreau de Loubresse, Charles, and Aegerter, Philippe (2007). "Natural history of progressive adult scoliosis". In: *Spine*. DOI: 10.1097/01.brs.0000263328.89135.a6.

Maturana, Daniel and Scherer, Sebastian (2015-09). "VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). DOI: 10.1109/IROS.2015.7353481.

Mazurowski, Maciej A., Dong, Haoyu, Gu, Hanxue, Yang, Jichen, Konz, Nicholas, and Zhang, Yixin (2023). "Segment anything model for medical image analysis: An experimental study". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2023.102918.

McKenna, Claire, Wade, Ros, Faria, Rita, Yang, Huiqin, Stirk, Lisa, Gummerson, Nigel W., Sculpher, Mark J., and Woolacott, Nerys (2012). "EOS 2D/3D X-ray imaging system: a systematic review and economic evaluation." In: *Health technology assessment*.

Meng, Di, Mohammed, Eslam, Boyer, Edmond, and Pujades, Sergi (2022). "Vertebrae Localization, Segmentation and Identification Using a Graph Optimization and an Anatomic Consistency Cycle". In: *Machine Learning in Medical Imaging (MLMI)*. DOI: 10.1007/978-3-031-21014-3_32.

Menze, Bjoern H., Jakab, Andras, Bauer, Stefan, Kalpathy-Cramer, Jayashree, Farahani, Keyvan, Kirby, Justin, Burren, Yuliya, Porz, Nicole, Slotboom, Johannes, Wiest, Roland, and Van Leemput, Koen (2015). "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". English. In: *IEEE Transactions on Medical Imaging*. DOI: 10.1109/TMI.2014.2377694.

Milletari, Fausto, Navab, Nassir, and Ahmadi, Seyed-Ahmad (2016). "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV 2016)*. Stanford, CA, USA: IEEE. DOI: 10.1109/3DV.2016.79.

Morid, Mohammad Amin, Borjali, Alireza, and Del Fiol, Guilherme (2021). "A scoping review of transfer learning research on medical image analysis using ImageNet". In: *Computers in Biology and Medicine*. DOI: 10.1016/j.compbiomed.2020.104115.

Morrissy, Raymond T., Goldsmith, Gregory S., Hall, Elmer C., Kehl, Douglas, and Cowie, G. Henry (1990). "Measurement of the Cobb angle on radiographs of patients who have scoliosis: Evaluation of intrinsic error". In: *The Journal of Bone and Joint Surgery A*. DOI: 10.2106/00004623-199072030-00005.

Myronenko, Andriy (2018). "3D MRI brain tumor segmentation using autoencoder regularization". In: *BrainLes Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. DOI: 10.1007/978-3-030-11726-9_28.

Ng, Phoebe Ting Ting, Straker, Leon, Tucker, K., Izatt, Maree, and Claus, Andrew (Mar. 2023). "Advancing Use of DEXA Scans to Quantitatively and Qualitatively Evaluate Lateral Spinal Curves, for Preliminary Identification of Adolescent Idiopathic Scoliosis". In: *Clinical and Translational Immunology (CTI)*.

Oktay, Ozan, Schlemper, Jo, Folgoc, Loic Le, Lee, Matthew, Heinrich, Mattias, Misawa, Kazunari, Mori, Kensaku, McDonagh, Steven, Hammerla, Nils Y., Kainz, Bernhard, Glocker, Ben, and Rueckert, Daniel (July 2018). "Attention U-Net: Learning Where to Look for the Pancreas". In: *Proceedings of the 1st Medical Imaging with Deep Learning (MIDL) Conference*. Presented at MIDL 2018; no formal proceedings volume. Amsterdam, The Netherlands. DOI: 10.48550/arXiv.1804.03999.

Oquab, Maxime, Bottou, Léon, Laptev, Ivan, and Sivic, Josef (2014). "Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pang, Sheng, Pang, Caizi, Su, Zhongyi, Lin, Li, Zhao, Lei, Chen, Yingwei, Zhou, Yue, Lu, Huimao, and Feng, Qianjin (2022). "DGMSNet: Spine segmentation for MR image by a detection-guided mixed-supervised segmentation network". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2021.102261.

Pasha, Saba (Nov. 2018). "Data-driven Classification of the 3D Spinal Curve in Adolescent Idiopathic Scoliosis with an Applications in Surgical Outcome Prediction". In: *Scientific Reports*. DOI: 10.1038/s41598-018-34261-6.

Peng, Songyou, Genova, Kyle, Jiang, Chiyu, Tagliasacchi, Andrea, Pollefeys, Marc, Funkhouser, Thomas, et al. (2023). "OpenScene: 3D Scene Understanding with Open

Vocabularies". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pfeffer, Maximilian Achim and Ling, Sai Ho (2022). "Evolving Optimised Convolutional Neural Networks for Lung Cancer Classification". In: *Signals*. DOI: 10.3390/signals3020018.

Pruijs, J. E., Hageman, M. A., Keessen, W., Meer, R. van der, and Wieringen, J. C. van (1994). "Variation in Cobb angle measurements in scoliosis". In: *Skeletal Radiology*. DOI: 10.1007/BF00223081.

Qadri, Syed Furqan, Lin, Hongxiang, Shen, Linlin, Ahmad, Masood, Qadri, Salman, Khan, Salabat, Khan, Maqbool, Zareen, Syeda Shamaila, Akbar, Muhammad Azeem, Bin Heyat, Md Belal, and Qamar, Shamsul (2023). "CT-Based Automatic Spine Segmentation Using Patch-Based Deep Learning". In: *International Journal of Intelligent Systems*. DOI: 10.1155/2023/2345835.

Qayyum, Adnan, Anwar, Syed Muhammad, Majid, Muhammad, Awais, M., and Alnowami, Majdi R. (2017). "Medical Image Analysis using Convolutional Neural Networks: A Review". In: *Journal of Medical Systems*.

Qi, Charles R, Su, Hao, Mo, Kaichun, and Guibas, Leonidas J (2017a). "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, Charles Ruizhongtai, Yi, Li, Su, Hao, and Guibas, Leonidas J (2017b). "Pointnet++: Deep hierarchical feature learning on point sets in a metric space". In: *Advances in neural information processing systems (NIPS)*.

Qu, Biao, Cao, Jianpeng, Qian, Chen, Wu, Jinyu, Lin, Jianzhong, Wang, Liansheng, Ou-Yang, Lin, Chen, Yongfa, Yan, Liyue, Hong, Qing, Zheng, Gaofeng, and Qu, Xiaobo (2022). "Current development and prospects of deep learning in spine image analysis: a literature review". In: *Quantitative Imaging in Medicine and Surgery*. DOI: 10.21037/qims-21-939.

Qu, Chongyu, Zhang, Tiezheng, Qiao, Hualin, Liu, Jie, Tang, Yucheng, Yuille, Alan L., and Zhou, Zongwei (2023). "AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks". In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al. (2021). "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning (ICML)*. Proceedings of Machine Learning Research. DOI: 10.48550/arXiv.2103.00020.

Rajpurkar, Pranav, Irvin, Jeremy, Bagul, Aarti, Ding, Daisy, Duan, Tony, Mehta, Hershel, Yang, Brandon, Zhu, Kaylie, Laird, Dillon, Ball, Robyn L., Langlotz, Curtis, Shpanskaya, Katie, Lungren, Matthew P., and Ng, Andrew Y. (2018). *MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs*. DOI: 10.48550/arXiv.1712.06957.

Ranjbarzadeh, Ramin, Kasgari, Abbas Bagherian, Ghoushchi, Saeid Jafarzadeh, Anari, Shokofeh, Naseri, Maryam, and Bendechache, Malika (2021). "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images". In: *Scientific Reports*. DOI: 10.1038/s41598-021-90428-8.

Rehm, Johannes, Germann, Thomas, Akbar, Michael, Pepke, Wojciech, Kauczor, Hans Ulrich, Weber, Marc-André, and Spira, Daniel (2017). "3D-modeling of the spine using EOS imaging system: Inter-reader reproducibility and reliability". In: *Public Library of Science (PLoS) ONE*. DOI: 10.1371/journal.pone.0171258.

Roaf, Robert (1958). "Rotation movements of the spine with special reference to scoliosis." In: *The Journal of bone and joint surgery. British volume*. DOI: 10.1302/0301-620X.40B2.312.

Rockenfeller, Robert and Müller, Andreas (Aug. 2022). "Augmenting the Cobb angle: Three-dimensional analysis of whole spine shapes using Bézier curves". In: *Computer Methods and Programs in Biomedicine*. DOI: 10.1016/j.cmpb.2022.107075.

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR52688.2022.01042.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer International Publishing. DOI: 10.1007/978-3-319-24574-4_28.

Roth, HR, Lu, L, Farag, A, Turkbey, E, Liu, F, and Summers, RM (2015). "Deep Convolutional Networks for Automated Vertebrae Detection and Segmentation in CT Images". In: *Medical Image Analysis*.

Ruppert, David, Wand, Matt P, and Carroll, Raymond J (2003). *Semiparametric regression*. Cambridge university press.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael S., Berg, Alexander C., and Li, Fei-Fei (2014). "ImageNet Large Scale Visual Recognition

Challenge". In: *Computing Research Repository (CoRR)*. DOI: 10.1007/s11263-015-0816-y.

Saeed, Muhammad Usman, Dikaios, Nikolaos, Dastgir, Aqsa, Ali, Ghulam, Hamid, Muhammad, and Hajjej, Fahima (2023). "An Automated Deep Learning Approach for Spine Segmentation and Vertebrae Recognition Using Computed Tomography Images". In: *Diagnostics*. DOI: 10.3390/diagnostics13162658.

Saito, Hiroshi and Tanaka, Yukio (2020). "Three-dimensional assessment of spinal deformity: Beyond the coronal plane". In: *Spine Journal*.

Sanders, James O., Khoury, Joseph G., Kishan, Surendra, and et al. (2008). "Predicting scoliosis progression from skeletal maturity: A simplified classification during adolescence". In: *The Journal of Bone and Joint Surgery A*. DOI: 10.2106/JBJS.G.00004.

Schuhmann, Christoph, Vencu, Richard, Beaumont, Romain, Kaczmarczyk, Robert, Mullis, Clayton, Katta, Aarush, Coombes, Theo, Jitsev, Jenia, and Komatsuzaki, Aran (2021). "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image–Text Pairs". In: *Proceedings of the NeurIPS 2021 Workshop on Data-Centric AI*. NeurIPS Workshop Paper.

Schwab, F., Dubey, A., Gamez, L., El Fegoun, A. B., Hwang, K., Pagala, M., and Farcy, J. P. (2005). "Adult scoliosis: prevalence, SF-36, and nutritional parameters in an elderly volunteer population". In: *Spine*. DOI: 10.1097/01.brs.0000160842.43482.cd.

Settles, Burr (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report. University of Wisconsin–Madison. URL: http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

Shao, Zhiwen, Yuan, Yichen, Ma, Lizhuang, Yeung, Dit-Yan, and Zhu, Xiaojia (2024). *SG-LRA: Self-Generating Automatic Scoliosis Cobb Angle Measurement with Low-Rank Approximation*. DOI: 10.48550/arXiv.2411.12604.

Shetty, Karthik, Birkhold, Annette, Jaganathan, Srikrishna, Strobel, Norbert, Egger, Bernhard, Kowarschik, Markus, and Maier, Andreas (2023). "BOSS: Bones, organs and skin shape model". In: *Computers in Biology and Medicine*.

Shi, Liangyu, Wang, Hongfei, and Shea, Graham Ka-Hon (2025). "The Application of Artificial Intelligence in Spine Surgery: A Scoping Review". In: *JAAOS Global Research and Reviews*. DOI: 10.5435/JAAOSGlobal-D-24-00405.

Shin, Hoo-Chang, Roth, Holger R., Gao, Mingchen, Lu, Le, Xu, Ziyue, Nogues, Isabella, Yao, Jianhua, Mollura, Daniel, and Summers, Ronald M. (2016). "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning". In: *IEEE Transactions on Medical Imaging*. DOI: 10.1109/TMI.2016.2528162.

Simons, Samuel, Wang, Songbo, Chen, Ximing, Wang, Yizheng, Choi, Euijoon, and Zhang, Li (2025). "SpineFM: Leveraging Foundation Models for Automatic Spine X-ray Segmentation". In: *arXiv preprint arXiv:2411.00326*. Submitted to IEEE ISBI 2025. URL: https://arxiv.org/abs/2411.00326.

Smailagic, Asim, Costa, Pedro, Noh, Hae Young, Walawalkar, Devesh, Khandelwal, Kartik, Galdran, Adrian, Mirshekari, Mostafa, Fagert, Jonathon, Xu, Susu, Zhang, Pei, and Campilho, Aurelio (2018). "MedAL: Accurate and Robust Deep Active Learning for Medical Image Analysis". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. DOI: 10.1109/ICMLA.2018.00078.

Smith, John et al. (2019). "Comparative analysis of adolescent and degenerative scoliosis using three-dimensional imaging". In: *Spine*.

Srinivas, A., Lin, Tsung-Yi, Parmar, Niki, Shlens, Jonathon, Abbeel, P., and Vaswani, Ashish (2021). "Bottleneck Transformers for Visual Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stokes, Ian AF, Armstrong, James G, and Moreland, Matthew S (2018). "Reducing radiation exposure in scoliosis imaging". In: *Spine Deformity*.

Su, Hang, Maji, Subhransu, Kalogerakis, Evangelos, and Learned-Miller, Erik (2015). "Multi-view convolutional neural networks for 3d shape recognition". In: *Proceedings of the IEEE international conference on computer vision*.

Sudlow, Cathie L. M., Gallacher, John E., Allen, Naomi E., Beral, Valerie, Burton, Paul, Danesh, John, Downey, Paul, Elliott, Paul, Green, Jane, Landray, Martin J, Liu, Bette C, Matthews, Paul M., Ong, Giok, Pell, Jill P., Silman, Alan J, Young, Alan, Sprosen, Tim, Peakman, Tim C, and Collins, Rory (2015). "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". In: *PLoS Medicine*. DOI: 10.1371/journal.pmed.1001779.

Tajbakhsh, Nima, Jeyaseelan, Laura, Li, Qian, Chiang, Jeffrey N, Wu, Zhihao, and Ding, Xiaowei (2020). "Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2020.101693.

Takmaz, Ayça, Fedele, Elisabetta, Sumner, Robert W., Pollefeys, Marc, Tombari, Federico, and Engelmann, Francis (2023). "OpenMask3D: Open-Vocabulary 3D Instance Segmentation". In: *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Tang, Samuel N., Walter, Anya, et al. (2024). "Applications of artificial intelligence for adolescent idiopathic scoliosis: mapping the evidence". In: *Spine Deformity*.

Taylor, H., Harding, I., Hutchinson, J., Nelson, I., Blom, A., Tobias, J., and Clark, E. (2013). "Identifying Scoliosis in Population-Based Cohorts: Development and Validation of a Novel

Method Based on Total-Body Dual-Energy X-Ray Absorptiometric Scans". In: *Calcified Tissue International*. DOI: 10.1007/s00223-013-9713-y.

Tribus, Clifford B. (2003). "Degenerative Lumbar Scoliosis: Evaluation and Management". In: *Journal of the American Academy of Orthopaedic Surgeons*. DOI: 10.5435/00124635-200305000-00004.

Van der Maaten, Laurens and Hinton, Geoffrey (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research*. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

Wang, Aobo, Zou, Congying, Yuan, Shuo, Fan, Ning, Du, Peng, Wang, Tianyi, and Zang, Lei (2024a). "Deep learning assisted segmentation of the lumbar intervertebral disc: a systematic review and meta-analysis". In: *Journal of Orthopaedic Surgery and Research*. DOI: 10.1186/s13018-024-05002-5.

Wang, Dequan, Shelhamer, Evan, Liu, Shaoteng, Olshausen, Bruno, and Darrell, Trevor (2021a). "Tent: Fully Test-Time Adaptation by Entropy Minimization". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=uXl3bZLkr3c.

Wang, Kaiping, Zhan, Bo, Zu, Chen, Wu, Xi, Zhou, Jiliu, Zhou, Luping, and Wang, Yan (2022). "Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2022.102447.

Wang, Liansheng, Xie, Cong, Lin, Yiping, Zhou, Hong-Yu, Chen, Kailin, Cheng, Dalong, Dubost, Florian, Collery, Benjamin, Khanal, Bidur, et al. (2021b). "Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior X-ray images: The AASCE2019 challenge". In: *Medical Image Analysis*. DOI: 10.1016/j.media.2021.102115.

Wang, W, Wang, Z, Zhu, Z, Zhu, F, and Qiu, Y (2016). "Body composition in males with adolescent idiopathic scoliosis: a case–control study with dual-energy X-ray absorptiometry". In: *BMC Musculoskeletal Disorders*. DOI: 10.1186/s12891-016-0968-0.

Wang, Wenxuan, Chen, Chen, Ding, Meng, Li, Jiangyun, Yu, Hong, and Zha, Sen (2021c). "TransBTS: Multimodal Brain Tumor Segmentation Using Transformer". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*. Lecture Notes in Computer Science.

Wang, Yihe, Huang, Nan, Li, Taida, Yan, Yujun, and Zhang, Xiang (2024b). "Medformer: A Multi-Granularity Patching Transformer for Medical Time-Series Classification". In: *Advances in Neural Information Processing Systems*.

Wasserthal, Jakob, Breit, Hanns-Christian, Meyer, Manfred T., Pradella, Maurice, Hinck, Daniel, Sauter, Alexander W., Heye, Tobias, Boll, Daniel T., Cyriac, Joshy, Yang, Shan, Bach, Michael, and Segeroth, Martin (2023). "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images". In: *Radiology: Artificial Intelligence*. DOI: 10.1148/ryai.230024.

Weinstein, Stuart L, Dolan, Lori A, Wright, James G, and Dobbs, Matthew B (2019). "Long-term outcomes of adolescent idiopathic scoliosis". In: *New England Journal of Medicine*.

Willemink, Martin J, Koszek, Wojciech A, Hardell, Cailin, Wu, Jie, Fleischmann, Dominik, Harvey, Hugh, Folio, Les R, Summers, Ronald M, Rubin, Daniel L, and Lungren, Matthew P (2020). "Preparing Medical Imaging Data for Machine Learning". In: *Radiology*. DOI: 10.1148/radiol.2020192224.

Windsor, Rhydian and Jamaludin, Amir (2020). "The Ladder Algorithm: Finding Repetitive Structures in Medical Images by Induction". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. DOI: 10.1109/ISBI45749.2020.9098469.

Windsor, Rhydian, Jamaludin, Amir, Kadir, Timor, and Zisserman, Andrew (2020). *A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI*. DOI: 10.1007/978-3-030-59725-2_69.

Windsor, Rhydian, Jamaludin, Amir, Kadir, Timor, and Zisserman, Andrew (2021). "Self-Supervised Multi-Modal Alignment for Whole Body Medical Imaging". In: *proceedings of Medical Image Computing and Computer Assisted Intervention*. DOI: 10.1007/978-3-030-87196-3_9.

Wu, Junde, Fang, Huihui, Zhang, Yu, Yang, Yehui, and Xu, Yanwu (2022). "MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model". In: *International Conference on Medical Imaging with Deep Learning*. DOI: 10.48550/arXiv.2211.00611.

Wu, Junde, Ji, Wei, Fu, Huazhu, Xu, Min, Jin, Yueming, and Xu, Yanwu (2024). "MedSegDiff-V2: diffusion-based medical image segmentation with transformer". In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'24/IAAI'24/EAAI'24. AAAI Press. DOI: 10.1609/aaai.v38i6.28418.

Yang, Jiancheng, Shi, Rui, Wei, Donglai, Liu, Zequan, Zhao, Lin, Ke, Bilian, Pfister, Hanspeter, and Ni, Bingbing (2023). "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification". In: *Scientific Data*. DOI: 10.1038/s41597-022-01721-8.

Yang, Lin, Zhang, Yizhe, Chen, Jianxu, Zhang, Siyuan, and Chen, Danny Z. (2017). "Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation". In: *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*. Springer. DOI: 10.1007/978-3-319-66179-7_46.

Ylikoski, M. and Tallroth, K. (1990). "Measurement variations in scoliotic angle, vertebral rotation, vertebral body height, and intervertebral disc space height". In: *Journal of Spinal Disorders*.

Zeng, Yihan, Jiang, Chenhan, Mao, Jiageng, Han, Jianhua, Ye, Chaoqiang, Huang, Qingqiu, Yeung, Dit-Yan, Yang, Zhen, Liang, Xiaodan, and Xu, Hang (2023). "CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Junhua, Lv, Liang, Shi, Xinling, Wang, Yuanyuan, Guo, Fei, Zhang, Yufeng, and Li, Hongjian (2013). "3-D reconstruction of the spine from biplanar radiographs based on contour matching using the hough transform". In: *IEEE IEEE Transactions on Biomedical Engineering (TBME)*.

Zhang, Ling, Wang, Xiaosong, Yang, Dong, Sanford, Thomas, Harmon, Stephanie A., Turkbey, Baris I, Wood, Bradford J., Roth, Holger R., Myronenko, Andriy, Xu, Daguang, and Xu, Ziyue (2020). "Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation". In: *IEEE Transactions on Medical Imaging*. DOI: 10.1109/TMI.2020.2973595.

Zhang, Renrui, Guo, Ziyu, Zhang, Wei, Li, Kunchang, Miao, Xupeng, Cui, Bin, Qiao, Yu, Gao, Peng, and Li, Hongsheng (2022). "PointCLIP: Point Cloud Understanding by CLIP". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR52688.2022.00836.

Zhang, Wei et al. (2019). "Limitations of 2D shape analysis in scoliosis: A critical review". In: *European Spine Journal*.

Zhang, Zhengxin, Liu, Qingjie, and Wang, Yunhong (2017). "Road Extraction by Deep Residual U-Net". In: *IEEE Geoscience and Remote Sensing Letters*.

Zhao, M, Meng, N, Cheung, Jason P Y, Yu, C, Lu, P, and Zhang, T (2023). "SpineHRformer: A Transformer-Based Deep Learning Model for Automatic Spine Deformity Assessment with Prospective Validation". In: *Bioengineering*. DOI: 10.3390/bioengineering10111333.

Zhou, Hong-Yu, Zhang, Xinru, Lu, Chixiang, Chen, Chaoqi, and Yu, Yizhou (2023). "nnFormer: Volumetric Medical Image Segmentation via 3D Transformer". In: *IEEE Transactions on Image Processing*. DOI: 10.1109/TIP.2023.3293771.

Zhou, Yujing, Liu, Yuan, Chen, Qian, Gu, Guohua, and Sui, Xiubao (2019). "Automatic Lumbar MRI Detection and Identification Based on Deep Learning". In: *Journal of Digital Imaging*. DOI: 10.1007/s10278-018-0130-7.

Zhou, Zongwei, Siddiquee, Md Mahfuzur Rahman, Tajbakhsh, Nima, and Liang, Jianming (2018). "UNet++: A Nested U-Net Architecture for Medical Image Segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...* DOI: 10.1007/978-3-030-00889-5_1.

Zhu, Jiayi, Bolsterlee, Bart, Chow, Brian V.Y., Song, Yang, and Meijering, Erik (2023). "Hybrid dual mean-teacher network with double-uncertainty guidance for semi-supervised segmentation of MRI scans". In: *arXiv preprint arXiv:2303.05126*. arXiv: 2303.05126.

Zhu, Xiaong (2005). *Semi-Supervised Learning Literature Survey*. Tech. rep. Computer Sciences Department, University of Wisconsin–Madison.

Zhu, Yuanpeng, Yin, Xiangjie, Chen, Zefu, Zhang, Haoran, Xu, Kexin, Zhang, Jianguo, and Wu, Nan (2024). "Deep Learning in Cobb Angle Automated Measurement on X-Rays: A Systematic Review and Meta-Analysis". In: *Spine Deformity*. Online ahead of print. DOI: 10.1007/s43390-024-00954-4.

Appendix A

Statement of Authorship

A statement of authorship is provided for each multi-authored paper included in this thesis. The statements describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication, there exists a complete statement that is filled out and signed by the candidate and supervisor.

Statement of Authorship for the paper "Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach" in Chapter 3.

Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach	Scoliosis Measurement on DXA Scans Using a Combined Deep Learning and Spinal Geometry Approach
Authors	Emmanuelle Bourigault, Amir Jamaludin, Timor Kadir, Andrew Zisserman
Publication status	Published
Publication details	Medical Imaging and Deep Learning, 2022.

Student Confirmation

Student name	Emmanuelle Bourigault	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models • writing and presentation of the paper	
Signature and Date		Apr. 24th 2025

Supervisor Confirmation

By signing the Statement of Authorship, the supervisor is certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name	Prof. Andrew Zisserman	
Supervisor comments		
Signature and Date		Apr. 24th 2025

Statement of Authorship for the paper "3D Shape Analysis of Scoliosis" in Chapter 4.

3D Shape Analysis of Scoliosis	3D Shape Analysis of Scoliosis
Authors	Emmanuelle Bourigault, Amir Jamaludin, Emma M.Clark, Jeremy Fairbank, Timor Kadir, Andrew Zisserman
Publication status	Published
Publication details	Shape MI, Medical Image Computing and Computer Assisted Intervention, 2023.

Student Confirmation

Student name	Emmanuelle Bourigault	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models • writing and presentation of the paper	
Signature and Date		Apr. 24th 2025

Supervisor Confirmation

By signing the Statement of Authorship, the supervisor is certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name	Prof. Andrew Zisserman	
Supervisor comments		
Signature and Date		Apr. 24th 2025

Statement of Authorship for the paper "3D Spine Shape from 2D DXA" in Chapter 5.

3D Spine Shape from 2D DXA	3D Spine Shape from 2D DXA
Authors	Emmanuelle Bourigault, Amir Jamaludin, and Andrew Zisserman
Publication status	Published
Publication details	Medical Image Computing and Computer Assisted Intervention, 2024.

Student Confirmation

Student name	Emmanuelle Bourigault	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models • writing and presentation of the paper	
Signature and Date		Apr. 24th 2025

Supervisor Confirmation

By signing the Statement of Authorship, the supervisor is certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name	Prof. Andrew Zisserman	
Supervisor comments		
Signature and Date		Apr. 24th 2025

Statement of Authorship for the paper "Automated DXA Scoliosis Method (DSM)" in Chapter 6.

Automated DXA Scoliosis Method (DSM)	Automated DXA Scoliosis Method (DSM)
Authors	Emmanuelle Bourigault, Amir Jamaludin, Emma M.Clark, Jeremy Fairbank, Timor Kadir, Andrew Zisserman
Publication status	Under submission to European Spine journal, 2025.
Publication details	

Student Confirmation

Student name	Emmanuelle Bourigault	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models • writing and presentation of the paper	
Signature and Date		Apr. 24th 2025

Supervisor Confirmation

By signing the Statement of Authorship, the supervisor is certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name	Prof. Andrew Zisserman	
Supervisor comments		
Signature and Date		Apr. 24th 2025

Statement of Authorship for the paper "UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation" in Chapter 7.

UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation	UKBOB: One Billion MRI Labeled Masks for Generalizable 3D Medical Image Segmentation
Authors	Emmanuelle Bourigault, Amir Jamaludin, Abdullah Hamdi
Published	
Publication details	International Conference on Computer Vision, 2025.

Student Confirmation

Student name	Emmanuelle Bourigault	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models • writing and presentation of the paper	
Signature and Date		Apr. 24th 2025

Supervisor Confirmation

By signing the Statement of Authorship, the supervisor is certifying that the candidate made a substantial contribution to the publication, and that the description above is accurate.

Supervisor name	Prof. Andrew Zisserman	
Supervisor comments		
Signature and Date		Apr. 24th 2025