Automated Radiological Grading of Spinal MRIs via Weak and Self-Supervision



Rhydian Llewellyn Windsor

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Abstract

The main aim of this thesis is to create automated tools that can interpret spinal MRI scans as accurately as an expert radiologist. This is motivated by the need to reduce waiting times for scan interpretation, address the issue of high variability among readers for common grading methods, and facilitate the scaling of clinical studies to larger cohorts. We begin by exploring automated vertebrae detection in clinical scans. First, we propose a sequential detection method, where the location of previous vertebra detection is used as a prior for the next. This is followed by an improved, language-modelling-inspired method that allows for rapid, accurate detection and labelling in arbitrary sequence and field-of-view. Next, we propose a novel transformer-based architecture for diagnosis of spinal disorders, demonstrating the importance of incorporating contextual information from surrounding vertebrae and multiple MR sequences in grading decisions. This is tested on disc degeneration and cancer datasets, yielding strong results in both cases.

Related prior research has relied on manually annotated data almost exclusively for model training. However, this has several limitations (e.g. significant time commitments and domain knowledge required from annotators, high inter-reader variability). Therefore, we set an auxiliary goal; to reduce the need for manual annotation in medical imaging through self-supervised learning. To this end, we propose a novel method for learning representations of same-subject, different-modality images pairs, using registration as a pretext task. This method is used to perform 3D spinal segmentation in MRIs using 2D DXA-scan annotations for supervision. Finally, we explore using free-text radiology reports as a source of weak supervision for representation learning. Experimenting with a dataset of chest X-rays, we adapt vision-language model training methods to improve performance when training data is limited, as is often the case in medical imaging.

Keywords – medical imaging, deep learning, spinal MRI, self-supervised learning.

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.

Rhydian Windsor, March 2023.

Acknowledgement

There are simply too many people I wish to thank and this section is already fairly long and self-indulgent. I have therefore split it into groups, in the hope that anyone I have not mentioned by name knows where they fit in.

Firstly, thanks to my funding institutions - I am deeply grateful to CRUK and the AIMS EPSRC CDT, for financial and administrative support thoughout my DPhil. In particular, to Wendy, who first welcomed me to Oxford with open arms and whose help and friendship has not wavered since.

To my supervisors and collaborators - Amir, I am amazed by the amount of time you have given me over the years. From fielding technical questions about research or software to indulging me while I grumble about reviewers, nothing ever seemed too much. Timor, thank you for bringing me to this project in the first place and for your enthusiasm and guidance throughout, whatever the challenge. Andrew, you have taught me more about being a good researcher than any other person and I will be forever grateful. Thanks also to our clinical team, in particular Sarim and Jeremy, for their time expended and knowledge imparted - this thesis would not be possible without them.

To my friends in VGG and AIMS, past and present (Henry, Prannay, Anna, Tom, Mark, Rob, Vitaly, Tim, Suny, Laurynas, Eldar, Tom, Luke, Tom, Sagar, Lili, Max and many more!) - Not only have I learned so much by being around so many talented people, but life has been made so much richer by the chats, laughs, coffees, runs and the occasional pint. Also, to my friends outside Oxford (George, Annie, the whole ex-Manchester crew, Louis, Matt, Olly) - you already know how much I appreciate your support and the last few years have been no exception.

To all my family - Thanks for your love and support. Mum, Dad & Bryn: any work ethic or intellectual curiosity I have is thanks to you and the example you set. Finally, to my fiancé, Lizzie. You've ridden this rollercoaster with me. You've been immeasurably kind and understanding in my worst times and celebrated my best times in equal measure. I love you and I can't wait for the next chapter of our lives together.

Contents

1 Introduction and Background **7**

1.1 Background and Key Ideas 9

1.2 Thesis Outline and Publications 17

I Automated Approaches To Diagnose and Monitor Spinal Disorders Using MRI **23**

2 The Ladder Algorithm **24**

2.1 Introduction 25

2.2 The Ladder Algorithm 27

2.3 Extracting Vertebral Bodies From Whole Spine MR Scans 28

2.4 Performance Evaluation & Results 31

2.5 Discussion & Conclusion 33

2.6 Acknowledgements 34

3 A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI **35**

3.1 Introduction 36

3.2 Detecting Vertebrae 38

3.3 Labelling Vertebrae 40

3.4 Detection and Labelling Results . . . 43
3.5 Automated Scoliosis Detection . . . 46
3.6 Conclusion . . . 47
3.7 Appendix . . . 48

4 Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading 51

4.1 Introduction . . . 53
4.2 Related Work . . . 54
4.3 Spinal Context Transformer . . . 55
4.4 Detecting Spine Cancer Using Radiological Reports . . . 57
4.5 Radiological Grading Of Spinal Degenerative Changes . . . 59
4.6 Experimental Results . . . 60
4.7 Conclusion . . . 62
4.8 Appendix . . . 63

II The Role of Self-Supervision In Radiological Imaging 68

5 Self-Supervised Multi-Modal Alignment For Whole Body Medical Imaging 69

5.1 Introduction . . . 71
5.2 Matching Scans Across Modalities . . . 74
5.3 Unsupervised Registration Of Multi-Modal Scans . . . 77
5.4 Conclusion . . . 81
5.5 Appendix . . . 82

6 3D Spinal Column Segmentation with Single Plane 2D-Projected Annotations 86

6.1 Introduction . . . 87
6.2 3D Segmentation with 2D Annotations . . . 88
6.3 Experiments . . . 90
6.4 Broader Impact Statement . . . 92
6.5 Acknowledgements . . . 93

7 Vision-Language Modelling for Radiological Imaging and Reports in the Low Data Regime **94**
7.1 Introduction . . . 96
7.2 Increasing Data Efficiency During Vision-Language Modelling . . . 98
7.3 Experiments . . . 101
7.4 Conclusion . . . 106
7.5 Appendix . . . 107

8 Discussion **117**
8.1 Impact . . . 117
8.2 Future Work . . . 119
8.3 Conclusion . . . 123

References **124**

A Statement of Authorship **140**

Chapter 1

Introduction and Background

Back pain is a widespread, pervasive condition, highly linked to spinal degeneration through ageing. The World Health Organisation estimates that it affects 1.71 billion people worldwide and is the leading cause of disability in over 80% of countries [World Health Organization 2022]. As the world's population grows and ages, this will only become more of an issue. Unfortunately, in many cases, current care is insufficient. For example, in the UK, referrals for spinal treatment grow year on year, and waiting lists become longer and longer. Accordingly, many patients miss the already-long 18-week target from referral to treatment [Greenough 2016]. This can lead to significant physical and psychological discomfort for the individual, as well as a huge socioeconomic burden on account of missed work and the additional care required while pain is ongoing.

Typically, 95% of back pain cases in the UK present to primary care (i.e. GP surgeries) as non-specific [National Institute for Health and Care Excellence (NICE) 2022]. In such cases, magnetic resonance imaging (MRI) is the gold-standard for finding clinical signs which could account for this pain since it has superior soft-tissue contrast compared to other modalities such as computed tomography (CT) and does not require exposure to harmful ionising radiation. It is also useful for ruling out potentially life-threatening or otherwise severely dehabilitating conditions such as spinal cancer, cord compression, bone fractures and infection. As MRI technology becomes cheaper and more ubiquitous, it presents an opportunity to significantly improve clinical care. However, this will introduce new challenges, particularly with regard to increasing radiologist workload. By definition, non-

specific back pain can occur for a wide range of reasons and therefore scans take time to read carefully. Furthermore, such conditions are rarely life-threatening and so radiologists may prioritise interpreting other, more urgent scans. In practice, this means it can be many days before a scan is read, slowing down patient care significantly.

Combined with recent advances in image analysis using deep learning, these factors make interpreting spinal MRI scans a promising candidate for automation. Not only would this reduce scan reading times and help alleviate growing backlogs[1], it would also increase reproducibility as different radiologists often interpret the same scan differently [Herzog et al. 2017]. Accordingly, this is the first goal of this thesis: to carve a path towards software capable of interpreting spinal MRIs in a manner akin to an expert radiologist. While a fully-automated general solution is the long-term goal, developing automated methods for a subset of radiologists' tasks also has many uses. For example, a system capable of quickly and accurately grading only one condition in a scan potentially enables much larger-scale studies into that condition where manual annotation of scans would otherwise bottleneck the number of possible study participants. As another example, a model which simply distinguishes between normal and abnormal scans could be used for automated triaging, allowing radiologists to attend to more urgent cases first.

Deep learning methods typically require large training datasets to properly account for the heterogeneity present in feature space. In natural images, this can be in terms of lighting, pose, occlusion, semantic content and so on. In medical imaging, this heterogeneity is often even more pronounced. For example, field of view can vary significantly between scans, as can the imaging technique used and the pathology of the subject being imaged. Therefore, perhaps the most significant challenge in developing such systems for spinal MRIs and other medical imaging investigations is a lack of large-scale, labelled, training datasets. There are several reasons for this: (1) In medical imaging, annotation usually requires a clinical expert whose time is both limited and expensive; (2) Considering the wide range of potential medical conditions pertinent to most investigations, as well as the complex relations between them, identifying the precise conditions to be annotated

[1]For more on this issue in the UK, see the Care Quality Commission Radiology Review (2018) https://www.cqc.org.uk/publications/themed-work/radiology-review.

is non-trivial; (3) sharing patient data between institutions and annotators is often difficult for legal and ethical reasons.

However, a comparatively easier task is to acquire large-scale, *unlabelled* datasets; these do not rely on the time and expertise of multiple experts to annotate, make no assumptions about the specific conditions present (since no fixed set of classes needs to be labelled), and data is shared between less parties, making ethical approval and patient consent easier to obtain. As such, a second goal of this thesis is to suggest methods of making progress in developing clinically useful models by leveraging such unlabelled datasets. This is primarily done through the use of *self-supervision*, whereby we learn a representation of data *without annotations*. In the following section, we describe self-supervision in more detail, along with several concepts key to the rest of this thesis.

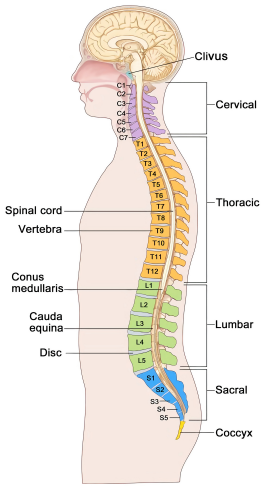
1.1 Background and Key Ideas

1.1.1 Spinal Anatomy

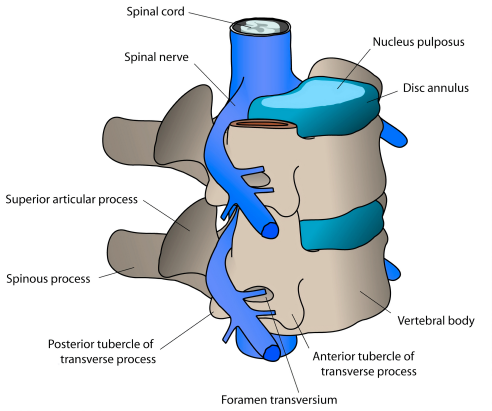
The human spine is a complex anatomical structure responsible for many physiological roles. It consists of multiple parts such as the spinal cord, spinal nerves, vertebrae, intervertebral discs as well as many muscles, ligaments and tendons. The components most important to the content of this thesis are shown in Figure 1.1 and described briefly below.

The *spinal cord* runs through the centre of the spinal column and is the key component of the central nervous system, transmitting sensory information between the brain and the rest of body. The cord is surrounded by *cerebrospinal fluid (CSF)*, which cushions the cord, protecting it from injury and providing nutrients. Together, the cord and CSF constitute the *spinal canal*.

Surrounding the spinal canal is the *vertebral column*. As well adding further protection to the spinal cord, the vertebrae are responsible for supporting the whole body while providing sufficient flexibility for motor function. The average human has 33 vertebrae, however this can vary between 32 and 35. A recent study estimated that these numerical variations occur in just under 8% of people [Tins and



(a) The human spine. Credit: Terese Winslow, NCI



(b) The vertebral column and spinal canal. Credit: New Heights Physical Therapy

Figure 1.1: Diagrams illustrating the anatomy of the human spine. (a) shows the five regions of the spine and their corresponding vertebrae. (b) shows how the vertebrae interlock with the spinal canal and intervertebral discs to form a complete column.

Balain 2016].

The vertebrae are typically categorised as belonging to one of five types depending on their location. From the bottom of the pelvis to the skull, these are the: *coccyx*, *sacrum*, *lumbar spine*, *thoracic spine* and *cervical spine*. The coccyx is 4 fused vertebrae at the bottom of the spine, creating the tailbone. The sacrum consists of 5 vertebrae (labelled S1-S5 from top to bottom), also fused together into a single bone attached to the pelvis. The lumbar spine (L1-L5) traverse the lower back. These are naturally curved inwards towards the anterior, known a *lordotic curve*. In around 3% of patients [Tins and Balain 2016], the bottom lumbar vertebra, L5, is fused to the top sacral vertebrae (S1), described as a *transitional vertebra*. The thoracic vertebrae (T1-T12) form the posterior portion of the rib cage. These are naturally curved towards the posterior, known as kyphosis. The spinal cord terminates at the bottom of the thoracic spine (between L1 and L2), becoming a bundle of nerves known as the *cauda equina*. The final part is the *cervical spine* (C1-C7), which runs from bottom of the skull to the top of the rib cage. These vertebrae also have slight lordosis. C1 is known as the *atlas*. This supports the skull and can articulate with C2, known as the *axis*, allowing for a wide range of head movement.

The *intervertebral discs* are cartilaginous structures located between each verte-

brae. These serve several functions, including load distribution, shock absorbtion and allowing for flexibility in the spine. The resulting gaps between vertebrae also allow for nerves to project from the spinal cord to the periphery of the body. Each disc consists of an tough exterior of ligament fibre, the *annulus fibrosus*, and a soft jelly-like interior, the *nucleus pulposus*.

1.1.2 Spinal Conditions

Since it is such a complex structure, there are a variety of conditions that can affect the spine's health and function. Some parts of the spine, in particular the intervertebral discs, can degenerate over time, and thus back pain is strongly correlated with age [Jamaludin et al. 2023]. This degeneration causes a wide range of issues, the most important and prevalent of which are described below. Of course, there are also other forms of spinal conditions caused by a myriad of reasons, e.g. acute trauma, infection or cancer metastasing to the spine. These are also briefly covered.

Disc Degeneration: Over time, intervertebral discs begin to wear down through repetitive stress. Since there is little blood flow to the discs, they have limited ability to repair themselves and therefore this is strongly associated with ageing, with around 70% of asymptomatic patients above the age of 40 having some form of degeneration [Brinjikji et al. 2015]. As a result, their ability to act as shock absorbers and also hold the *endplates* of the adjacent vertebrae apart lessens. These bones will often rub against each other, leading to pain and stiffness through *osteoarthritis*. The narrowing can also compress the spinal canal, in a condition known as *spinal stenosis*. Occasionally, either through degradation or trauma, the disc nucleus can *herniate* through the annulus ligament that holds it in place. Since the spinal column is fairly compact, again compressing the spinal cord and often causing severe pain. A widely-used method to assess the degree of disc degeneration is *Pfirrmann grading* [Pfirrmann et al. 2001]. Taking into several factors such as disc height, structure, signal intensity and the distinction between the annulus and nucleus, this system assigns a grade from 1 (healthy disc) to 5 (severe disc degeneration).

Bone Degeneration: Like all types of bone, vertebrae also degenerate over time. As bone mass decreases, they become more brittle and the risk of fractures and related conditions increases. In vertebrae, hairline fractures can occur at the vertebral endplates which grow over time, leading to defects. Furthermore, vertebrae can slip backwards or forwards, out of alignment with the rest of the spine in a process known as *spondylolisthesis*. This decreases spinal stability and risks cord compression. Another common phenomenon is *modic changes*, whereby the signal intensity in vertebral endplates changes, becoming either hypo- or hyper-intense in some regions. The exact nature of these lesions is not well-understood, however they have been strongly linked with lower back-pain [Kjaer et al. 2005].

Scoliosis: Scoliosis is a term for describing a sideways curve of the spine (i.e. from left to right). This can be caused by a number of factors, and in most cases the exact cause is unknown (*idiopathic scoliosis*). This can cause back pain, breathing problems, as well as body image issues. The most widely used method to assess the severity of scoliosis is the *Cobb angle*. This refers to the angle between the lines drawn parallel to the endplates of the vertebrae at the top and bottom of the curve. Normally, a Cobb angle of over 10° is regarded as scoliosis, with severe cases having angles of over 20° and surgery usually required for angles of greeater than 45° [Eisenberg and Weinstein 2022].

Spinal Cancer: While primary cancers of the spine are fairly rare, cancer can play a major role in the development of spinal conditions. One common issue is *spinal metastases*, whereby cancer cells spread from their primary tumour to the spine. This in relatively common; 2 in 3 patients with cancer will develop some form of metastasis [Shaw et al. 1989], with the spine being the most common site for this to occur [Ortiz Gomez 1995]. While patients are often palliative by the time they have multiple spinal metastases, they can have a severely deleterious effect on the patient's well-being. Typically, this is due to the risk of vertebral fractures and also spinal cord compression. Regular monitoring of metastases is therefore important, allowing for early intervention if either of these complications are about to occur. Another cancer with a strong association with the spine is *multiple myeloma*. This is a form of bone marrow cancer which can severely reduce

bone strength. Again, the risk of vertebral fractures increases significantly, which in turn can limit mobility, cause pain and lead to cord compression.

1.1.3 Spinal MRIs

Magnetic resonance imaging (MRI) is a powerful diagnostic imaging technique. Images are created by measuring the current induced by the changing net magnetisation of excited hydrogen nuclei over time as they relax from a higher to a lower energy state, moving into alignment with a strong external magnetic field. The characteristics of the resulting current (the *MR signal*) gives information on the local molecular environment of the protons which in turn can be used to infer the type of tissue. The image depends on the timing by which hydrogen nuclei are excited by radiofrequency electromagnetic waves, as well as the gradient of the magnetic field applied. These parameters are collectively known as the *pulse sequence* (often shortened to just *sequence*). Employing different pulse sequences allows for the emphasis of varying tissue types depending on the clinical investigation being conducted. In addition, patients can be dosed with substances which enhance certain types of tissue by binding to the hydrogen nuclei and altering their local environment (e.g. gadolinium).

To reduce the time taken to acquire the image, spinal MRIs are typically sliced - meaning resolution is lower in one axis than the other two (for example, voxels might represent a region of volume $0.5 \times 0.5 \times 10$ mm). This lower-resolution axis defines the plane of the scan which can be *axial, coronal* or *sagittal*, as shown in Figure 1.2. In clinical practice, depending on the investigation being conducted, usually only part of the spine is imaged, e.g. the lumbar portion. Occasionally, multiple images of different parts of the spine are registered together, creating a single combined whole-spine image in a process known as *stitching*. Common sequences used in spinal MRIs are T1, T2 and STIR. These serve different purposes, emphasising bony structures (T1), soft tissue (T2) and suppressing signal from fatty tissues (STIR). An example study of sagittal lumbar MRIs is shown in Figure 1.3.

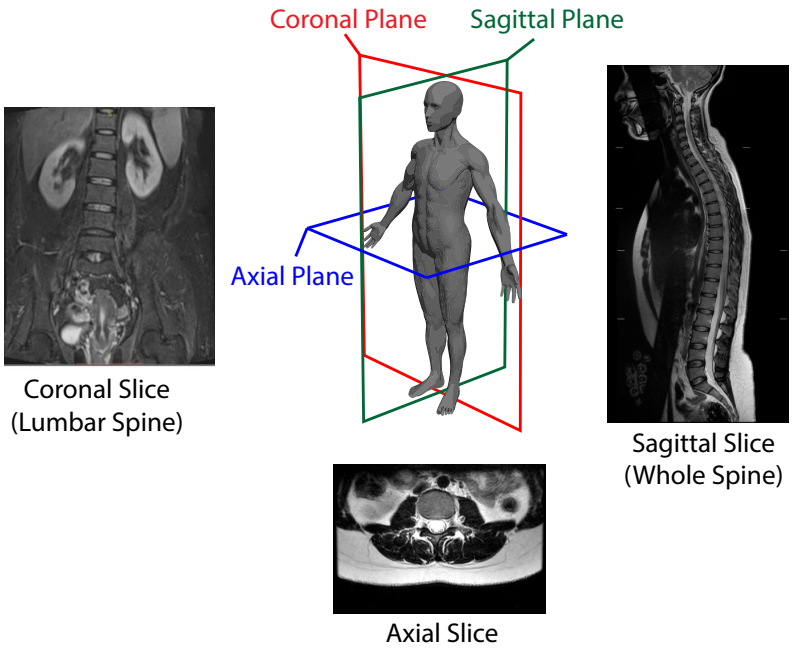


Figure 1.2: The axial, coronal and sagittal planes of a spinal MRI. In each case a slice of a typical MR scan in that plane is shown. Adapted version of a figure originally from [Ghoshal et al. 2019].

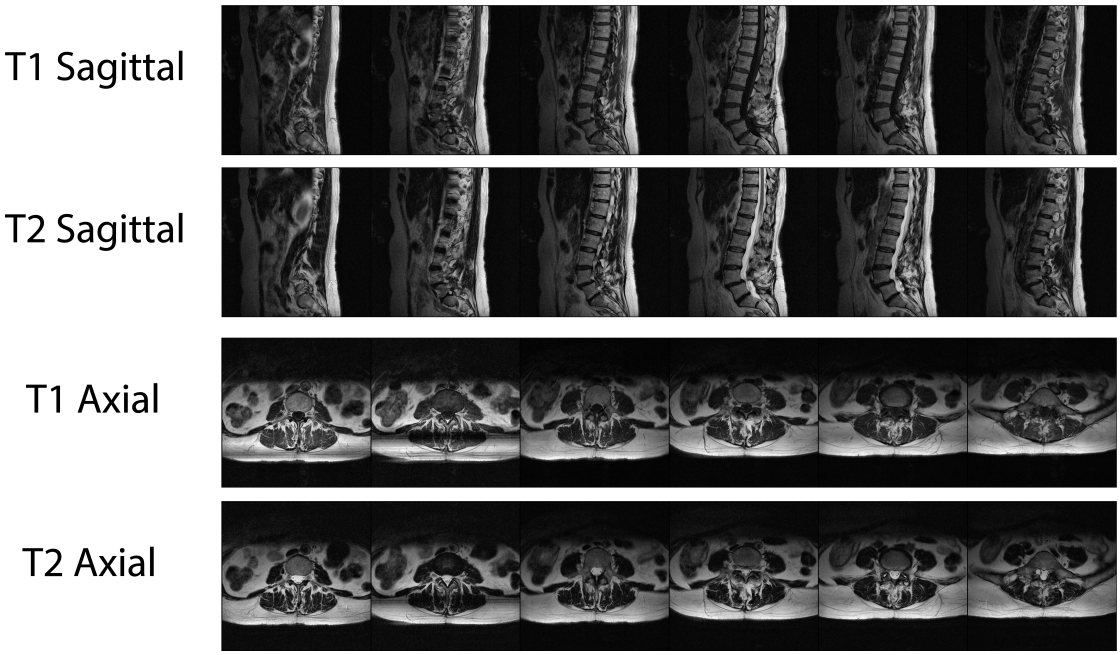


Figure 1.3: Example T1 & T2 study of sagittal lumbar MRIs from the Genodisc dataset. Note that only a subset of all possible slices are shown. Note the increased intensity of the vertebral discs and spinal cord (directly behind the vertebrae) in the sagittal and axial T2 images compared to T1.

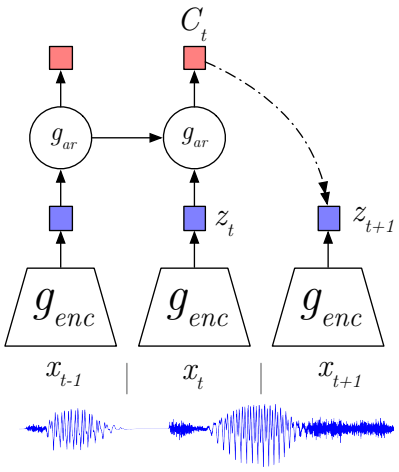
1.1.4 Self-Supervision & Weakly-Supervised Learning

Self-supervision refers to a paradigm in machine learning whereby a model learns to perform a task by exploiting the structure and relationships within the input data itself, rather than relying on externally provided labels or annotations (as in classically supervised learning). This approach allows the model to generate its own supervision signals using the inherent structure of the data, effectively turning unsupervised learning problems into supervised ones. For example, in computer vision, one might mask out part of an image and task a model to predict the missing pixels using the context provided by the rest of the image. Some examples of common self-supervised tasks across a range of modalities (e.g. images, audio and text) are shown in Figure 1.4.

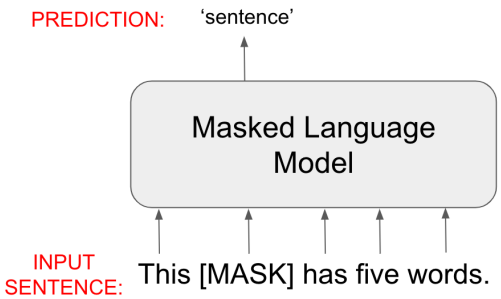
This approach has become increasingly popular in recent years, largely due to the ease of acquiring large amounts of unlabeled data compared to high-quality, labelled data. Accordingly, models can be trained on a much larger scale in terms of data which, coupled with increased availabilty of compute resources, has led to significant improvements in performance. Recent high-profile examples include the success of large language models such as GPT-3 [Brown et al. 2020] and BERT [Devlin et al. 2019] in natural language processing and approaches such as SimCLR [T. Chen et al. 2020], MoCo [He et al. 2020], DINO [Caron et al. 2021] and MAE [He et al. 2021] in computer vision.

The success of self-supervised learning is particularly relevant to medical imaging. Primarily, this is because annotating medical images typically requires a trained medical professional, whose time is limited and expensive. Furthermore, annotating each sample can be very time-consuming for some tasks (e.g. creating 3D segmentation maps). Furthermore, sharing the annotation burden among multiple labellers can introduce challenges related to significant inter-annotator variability, as well as the necessity to maintain patient privacy while exchanging data. Together, these factors strongly limit the size of possible annotated training datasets.

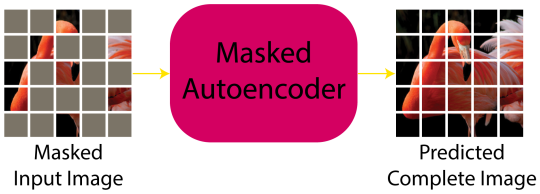
Self-supervision provides a partial solution to this problem by allowing a model to learn a good representation of the data from a large amount of unlabelled data which can, with appropriate anonymisation, consent and ethical clearance, quickly be extracted in bulk from hospital databases. This representation can then be fine-



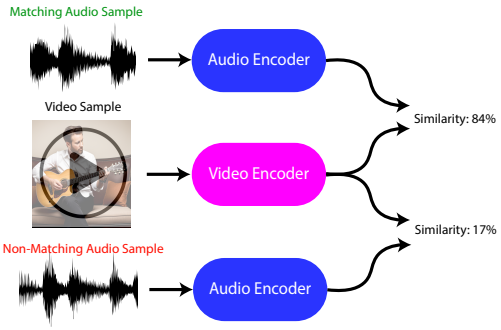
(a) Next sample prediction in audio [Oord et al. 2018]



(b) Masked word/token prediction [Devlin et al. 2019]



(c) Image Inpainting [He et al. 2021]



(d) Audiovisual Contrastive Learning [Arandjelović and Zisserman 2018]

Figure 1.4: Examples of self-supervised learning tasks across a variety of domains. (a) shows an example of a self-supervision method for audio, where the model is tasked with predicting the next sample in a sequence using previous samples for context. (b) shows *masked language modelling*, common pretraining task for natural language processing where a model must predict a token based on surrounding words. (c) is an example of self-supervision for images; the model must fill in the masked regions using context from the remaining patches in the image. Finally (d) shows an example of *multimodal self-supervision*; in this case, the model must match together the corresponding audio and frames from an existing video.

tuned on a smaller bespoke labelled dataset relevant to the task at hand.

A related idea is that of *weak supervision*, which is a more general approach to learning from limited supervision. In this approach, we have access to some form of supervisory signal, however it is assumed to be imperfect. This could be through noisy, missing or incomplete labels or through cases where labels exist for sets of samples rather than each individually (*multiple instance learning*). A concrete example of weak supervision in medical imaging is the use of radiologist reports to extract labels for image classification tasks - these reports will often not make reference to the condition of interest or suggest a wide range of possible explanations for the image. However, despite these drawbacks, weak supervision is often much more scalable than full supervision. For example, radiology reports can be extracted in bulk from hospital databases and thus avoid the need for retrospective annotation of potentially thousands of scans.

1.2 Thesis Outline and Publications

Chapters 2 to 7 contain papers published at biomedical imaging conferences over the course of my DPhil. Accordingly, these have all been subject to a peer-review process. There has been no substantial change to these papers, except in formatting. Statements of authorship outlining my contributions to each is given in Appendix A. The papers are introduced and briefly described below in the context of the ideas presented in Sections 1 and 1.1:

1.2.1 Part I: Back Pain Diagnosis and Monitoring

Chapter 2: "The Ladder Algorithm: Finding Repetitive Structures in Medical Images by Induction", **Rhydian Windsor**, Amir Jamaludin. *International Symposium on Biomedical Imaging (ISBI) 2020.* This paper presents a method for detecting vertebrae in whole spine MRIs using only lumbar MRIs as training data. This is a vital prerequisite task for models operating on spinal MRI scans, e.g. for automated spinal disease diagnosis. The method is inspired by inductive proofs in mathematics, in that the detection of the n-th vertebra is used to give a prior for

the location of the $(n+1)$-th vertebra. In this manner, the algorithm is reminiscent of climbing the rungs of a ladder, hence the name.

Chapter 3: "A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI", **Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020*. Like the previous chapter, this paper presents a method for detecting and labelling vertebrae, in this case in arbitrary fields of view and a wide range of MR sequences. The key improvement over the previous chapter is that this method does not require an 'anchor' vertebra to start from. This is achieved by the introduction of *Vector Field Regression (VFR)*, a method for detecting vertebrae whereby a series of landmarks are located and then grouped into vertebral instances by learned 'pointing' vector fields. In addition to not requiring an anchor, we show empirically that VFR is faster and considerably more robust to missed detections than the Ladder Algorithm. Finally, we illustrate that the detections of the resulting model can be used for the automatic diagnosis of scoliosis, a common spinal deformity.

Chapter 4: "Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading", **Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022*. In this chapter, focus is shifted to automated diagnosis of spinal disorders. We propose the Spinal Context Transformer (SCT), a new architecture for performing radiological grading of intervertebral discs. The key insight behind this architecture is that additional context is provided by viewing neighbouring IVDs and other MR sequences, and this can be used to inform grading predictions. Training this model on a dataset of lumbar MRIs annotated by an expert radiologist, we show that the SCT outperforms the original SpineNetV2 grading network. Additionally, and perhaps more importantly, SCT can take a broader range of MR sequences as input without adaption, and still achieve good performance. We also train SCT to detect spinal metastases, a common complication of many cancers, and the related conditions of vertebral compression fractures and cord compression. The final model shows good agreement with expert radiologists on a dataset of spinal MRIs, extracted from the local hospital trust.

1.2.2 Part II: Self-Supervision In Radiological Imaging

Chapter 5: "Self-Supervised Multi-Modal Alignment for Whole Body Medical Imaging", **Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021.* This work is motivated by a desire to reduce the amount of labelled data needed to train medical imaging models, by learning representations of data in a self-supervised manner. We consider cases where images of the same subject are available in multiple imaging modalities, a relatively common scenario in medical imaging. The paper introduces a new method of self-supervised learning for such occasions, the key idea of which is to contrastively match together *spatial* representations of images from different modalities into a joint embedding space. Applying this to a dataset of 20,000 paired Dixon MRI and DXA studies from the UK Biobank, our method shows incredibly strong cross-modal retrieval performance. Furthermore, we show the representations learned are useful for a range of downstream tasks, including registration and cross-modal segmentation.

Chapter 6: "3D Spinal Column Segmentation with Single Plane 2D-Projected Annotations", **Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman. *MedNeurips 2021.* Expanding on the work presented in Chapter 5, here we present a method to segment the spinal column in 3D MRIs using only 2D segmentations of the spine in DXA scans as training data. We suggest several improvements to a naive baseline, such as encoraging symmetric filters and using our prior knowledge of the spinal anatomy to design a loss function which encourages segmentations which are sparse in the coronal plane.

Chapter 7: "Vision-Language Modelling for Radiological Imaging and Reports in the Low Data Regime", **Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman. *Medical Imaging with Deep Learning (MIDL) 2023* (Oral presentation). Here, we explore methods of adapting popular vision-language models (VLMs) to the radiological domain. Specifically, in order to reduce the reliance of medical imaging models on large amounts of labelled data, we investigate the use of radiological reports as a source of supervision. Since medical imaging datasets are typically smaller than the millions of image-text pairs used to train conven-

tional VLMs, we are particularly interested in methods of training such models can be done with less data than typically used. Using a publically available set of chest X-rays and associated reports, we empirically test several methods to do this. Our final recommendations result in similar text-to-image retrieval performance compared to CLIP with 10 times less data. We also show that this holds for the downstream task of image classification.

Finally, Chapter 8 concludes the thesis with a discussion of the impact of the work presented, as well as potential future research direction.

1.2.3 Work Not Included

The following are other works completed during my DPhil that are not included in this thesis. The content is briefly described, along with a link to the paper if applicable, and reasons for not including it in the thesis.

Journal Papers:

Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A. (2023) "Automated Detection, Labelling and Radiological Grading of Clinical Spinal MRIs." Under Review, Scientific Reports - This paper describes SpineNetV2 from a methodological perspective, and evaluates its performance on several clinical datasets. This is not included since the majority of the technical content is covered in Chapters 3 and 5.

McSweeney, T. P., Tiulpin, A., Saarakkala, S., Niinimäki, J., **Windsor, R.**, Jamaludin, A., Kadir, T., Karppinen, J. and Määttä, J. (2022). "External Validation of SpineNet, an Open-source Deep Learning Model for Grading Lumbar Disc Degeneration MRI Features, Using the Northern Finland Birth Cohort 1966." Spine, 10.1097. - This constitues a validation study of SpineNet, tested on a dataset of lumbar MRIs from Finnish hospitals. The study concludes SpineNet matches human reliability and demonstrates robust performance despite the multiple challenges facing model generalizability. This work does not contribute any new technical content, and is therefore not included. Futhermore my role is the paper was limited to technical guidance on running the software implementation of SpineNet.

Conference Papers:

"Contouring by Unit Vector Field Regression", Amir Jamaludin, Sarim Ather, Timor Kadir, **Rhydian Windsor**, *International Symposium on Biomedical Imaging (ISBI) 2023*. This work presents a novel method for delineating open contours in medical images, relying on 'walking' along learnt vector fields. This concept is inspired by the vector field regression detection method presented in Chapter 3. We apply the method to delineating sacro-illiac joints in MRI scans of the pelvis, an important preliminary task when assessing the progress of ankylosing spondylitis inflammation. The work is not included in the thesis since it has little do with either spinal MRIs or self-supervised learning, the main foci of this thesis.

Abstracts:

"Machine Learning Based Berlin Scoring Of Magnetic Resonance Images of the Spine in Patient wth Ankylosing Spondylitis from the Measure 1 Study", Amir Jamaludin, **Rhydian Windsor**, Sarim Ather, Timor Kadir, Andrew Zisserman, Juergen Braun, Lianne Gensler, Pedro Machado, Mikkel Ostergaard, Denis Poddubnyy, Thibaud Coroller, Brian Porter, Shephard Mpofu, Aimee Readie. *European Congress of Rheumatology (EULAR) 2020.* - This abstract describes a system based on SpineNetV2 for the automated grading of ankylosing spondylitis progression in spinal MRIs.

"Machine Learning-based Berlin Scoring of Magnetic Resonance Images of the Spine in Patients with Ankylosing Spondylitis: Analysis of Data from a Phase 3 Trial with Secukinumab", Amir Jamaludin, **Rhydian Windsor**, Sarim Ather, Timor Kadir, Andrew Zisserman, Juergen Braun, Lianne Gensler, Pedro Machado, Mikkel Ostergaard, Denis Poddubnyy, Thibaud Coroller, Brian Porter, Shephard Mpofu, Aimee Readie. *American College of Rheumatology Convergence (ACR) 2020* - This describes a follow-up to the previous abstract with improved results.

"A Novel Deep-Learning Architecture to Detect Spinal Metastases, Vertebral Fractures and Cord Compression using Information Extracted from Radiological Reports", Sarim Ather, **Rhydian Windsor**, Sandeep Singh, Rajat Chowdhury, Amir Jamaludin, Jeremy Fairbank. *Royal College of Radiologists Annual Conference*

(RCR) 2022. This is a clinicaly-focused follow-up to the work presented in Chapter 4. It describes further analysis of SCT applied to spinal metastases detection, with comparision to additional expert radiologists.

Miscellaneous:

Windsor, R., Jamaludin, A., Kadir, T. and Zisserman, A. (2022). "SpineNetV2: automated detection, labelling and radiological grading of clinical MR scans" Technical Report, arXiv:2205.01683. - This report describes the initial SpineNetV2 software release[2] from a technical perspective, providing reference material for users of the package.

[2] https://github.com/rwindsor1/SpineNet

Part I

Automated Approaches To Diagnose and Monitor Spinal Disorders Using MRI

Chapter 2

The Ladder Algorithm

This chapter presents a paper published at ISBI 2020, the 17$^{\text{th}}$ International Symposium on Biomedical Imaging (Iowa/Virtual).

The Ladder Algorithm: Finding Repetitive Structures in Medical Images by Induction

Rhydian Windsor, Amir Jamaludin

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford

Abstract

In this paper we introduce the Ladder Algorithm; a novel recurrent algorithm to detect repetitive structures in natural images with high accuracy using little training data.

We then demonstrate the algorithm on the task of extracting vertebrae from whole spine magnetic resonance scans with only lumbar MR scans for training data. It is shown to achieve high performance with 99.8% precision and recall, exceeding current state of the art approaches for lumbar vertebrae detection in T1 and T2 weighted scans. It also generalises without retraining to whole spine images with minimal drop in accuracy, achieving 99.4% detection rate.

2.1 Introduction

Labelling medical images is time-consuming, labour-intensive and boring. This is especially true when detecting or segmenting repetitive structures such as vertebrae, teeth and ribs. Fortunately, we can leverage prior knowledge of spatial relations between instances of repeating objects to perform this task in an automated manner [Leung and Malik 1996; Hays et al. 2006]. In this paper we explore a new method of using a priori knowledge of structures such as these to ease the task of detection, resulting in a learning system needing fewer training examples to achieve good performance.

We demonstrate this algorithm on the task of detecting and labelling vertebral

bodies in whole spine magnetic resonance (MR) scans. Extracting vertebrae is an important intermediate task for the automated analysis of spinal diseases and there are several examples of this task being achieved with high accuracy in lumbar-specific MR scans [Lootus et al. 2014b; Jamaludin et al. 2015; Jamaludin et al. 2017c; Forsberg et al. 2017; Lu et al. 2018]. However, there is very little recent work on extracting vertebrae from whole spine scans. Such scans are important for diagnosing whole spine diseases such as ankylosing spondylitis, bone metastases and multiple myeloma. Most previous approaches for vertebral extraction in the whole spine case do not use deep learning, relying instead on fitting a polynomial to intensity profiles of scans and finding local minima to separate vertebrae [Peng et al. 2005] or random forest classifiers applied to appearance based features [Glocker et al. 2013].

Perhaps the reason that deep learning systems have not been popular for MR whole spine vertebral body extraction is a deficiency of labelled data. This may be because many MR scans are lumbar-only as this is the most common site for back pain to occur. Also, whole spine scans take much longer than lumbar scans to annotate, so generating datasets of a similar scale to those available for lumbar scans is difficult.

As a solution to this problem, we train a model to extract vertebral bodies from lumbar scans in such a way that it will generalise to whole spine scans. To achieve this, we introduce the Ladder Algorithm; starting at the S1 vertebra at the bottom of the spine each vertebra is bounded iteratively, using the previously detected vertebral body to infer its approximate location. This model is trained using a large cohort of labelled lumbar scans and then applied to whole spine scans simply by increasing the number of iterative predictions made. A single iteration of the algorithm is shown in Figure 2.1.

We evaluate the algorithm by assessing performance on manually labelled 2-D mid-sagittal slices from lumbar and whole spine scans. Following evaluation heuristics previously used in automated vertebral extraction, in particular those of [Lootus et al. 2014b; Forsberg et al. 2017], the algorithm achieves higher accuracy than state of the art approaches at detecting vertebral bodies in lumbar images. More importantly, the algorithm only has a small drop in accuracy when generalising to whole spine scans and as such is a cheap and data-efficient way to annotate them.

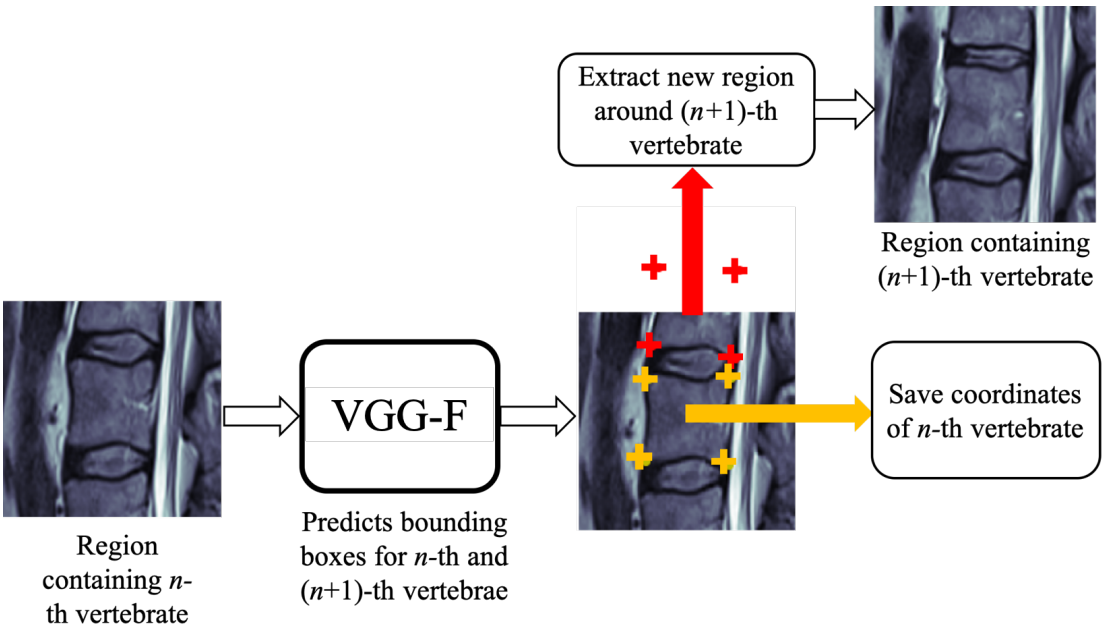


Figure 2.1: A single iteration of the Ladder Algorithm applied to vertebral body extraction. Given a image patch containing the n-th instance of a repetitive object, a network predicts a bounding box around the object (shown by yellow crosses) and proposes a location for the $(n+1)$-th instance (shown by red crosses). This proposal is used to extract the next patch.

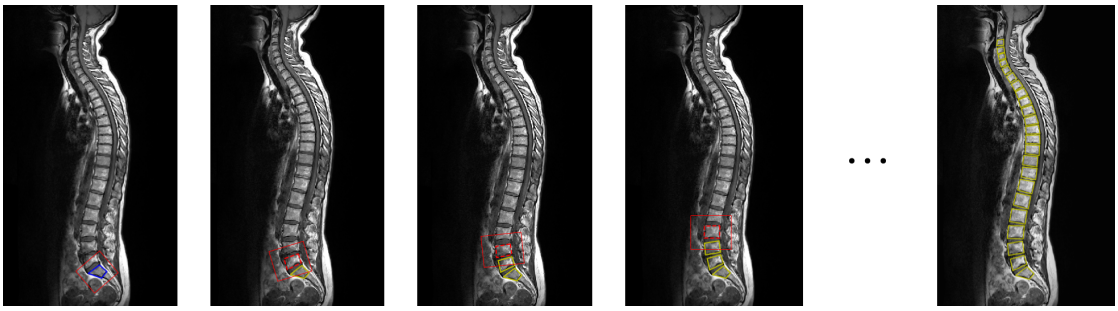


Figure 2.2: A demonstration of the steps used to detect vertebral bodies for the whole spine. Vertebral detections are made sequentially for a fixed number of iterations up the spine. Here the blue bounding box indicates the coordinates used to initialise the algorithm, the dashed red box indicated proposed regions containing vertebrae and the continuous red box is the patch given as input to the CNN. Yellow boxes indicate saved detections.

2.2 The Ladder Algorithm

The ladder algorithm is analogous to proof by induction. In inductive proof, the key step is to show that if some property holds for the n-th case of a sequence, it must also hold for the $(n+1)$-th case using knowledge of how the sequence was

generated. In the ladder algorithm, given a patch of an image containing the n-th instance of some repetitive object we use a convolutional neural network to predict a region containing the location of the $(n+1)$-th instance of the object as well as detecting the n-th object from the patch. If we expect N total instances of the object and can accurately extract the $n = 1$ case, then detecting all the objects is as simple as applying the algorithm for N iterations. Alternatively, if we do not know how many instances of an object exist in an image then the algorithm can be augmented to include a stopping criterion, identifying when we reach the final object in a sequence and stopping the algorithm.

Detecting objects as a sequence where each prediction is conditioned on previous predictions also has parallels to recurrent neural networks (RNNs) in the case of mapping a single input onto a sequence as an output. However, instead of using a network's hidden state to inform the next prediction we are more explicit, giving a specific patch of the image for the network to use to make its next prediction. The idea of using induction to train RNNs has been explored previously and shown to yield good results at tasks such as text spotting, counting and segmenting repeating objects [Gupta et al. 2018; Romera-Paredes and Torr 2016].

Having introduced the algorithm, the rest of this paper focuses on its application to the extraction of vertebral bodies from spinal MR scans. However, the algorithm can be applied to much wider range of domains. We suggest some potential medical domains for future work in Section 2.5.

2.3 Extracting Vertebral Bodies From Whole Spine MR Scans

In the experiments in this paper, the ladder algorithm is used to detect bounding boxes around vertebrae in whole spine scans using only vertebrae from lumbar scans as training data. Specifically, we attempt to predict from the S1 vertebra to the C3 vertebra in whole spine scans using annotated sequences of S1 to T12 vertebral bodies in lumbar spine scans as training examples. We do not attempt to detect the C1 and C2 vertebral bodies as they appear very different to the rest of the vertebrae in the spine.

An example of the algorithm being used to localise vertebral bodies in a whole spine scan is shown in Figure 6.1. Firstly, a detector trained on extracting vertebrae from lumbar MR scans such as [Lootus et al. 2014b] is used to extract the location of the S1 vertebra from a whole spine scan. This bounding box is used to extract a patch surrounding the S1 vertebral body which is then input to the network shown in Figure 2.1. This predicts co-ordinates in the patch for vertices of bounding boxes for the S1 vertebra and the vertebra above it, L5 (details on how the network is trained are given in Section 2.3.2). The resulting image coordinates of S1 are saved as final predictions whereas the coordinates of the L5 are used to extract a new patch of the image, centred on these coordinates. This new patch is then used as a new input to the network, which extracts the final coordinates of L5 and proposed coordinates for the body of the next vertebra, L4. This process repeats iteratively up the spine for a fixed number iterations. In the experiments in this paper, 23 iterations are used for full spine scans, resulting in detection of all vertebrae from S1 to C3 in almost all cases. However, an estimated 7.7% of the population have variations in the number of vertebrae in the spine [Tins and Balain 2016], meaning the exact range of vertebrae found cannot be determined trivially. The impact of this is discussed in Section 2.5.

2.3.1 Datasets Used

As stated, two datasets are used for the experiments in this paper. **Genodisc** is a large dataset consisting of 5404 T1 and T2 weighted MR lumbar scans from 2295 unique patients collected by the Genodisc consortium. Each scan has ground truth bounding boxes for S1, L1-5 and T12, extracted by the method outlined by Lootus *et al.* [Lootus et al. 2014b] and manually checked for accuracy. In this experiment, Genodisc is randomly split into training, validation and test sets with splits of 80%, 10% and 10% respectively. Splits were done to ensure that all scans of each patient remained in the same set.

Oxford Whole Spine (**OWS**) is a dataset of 64 T1 and T2 weighted 3-D clinical whole spine MR scans from unique patients. The cohort consists of a range of degenerative changes including scoliosis, stenosis and collapsed vertebrae. These scans are manually annotated with bounding boxes around each of the 23 vertebrae

from S1 to C3. OWS is used here solely as a testing dataset.

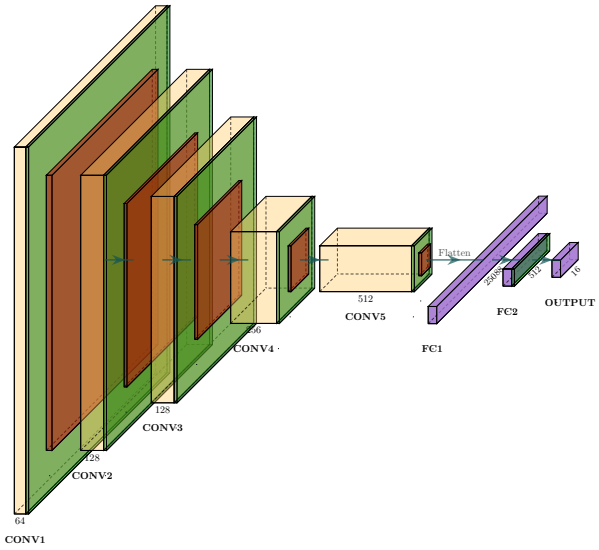


Figure 2.3: The architecture used recurrently to extract vertebral corners. Here yellow blocks are 2-D convolutional layers, green blocks represent batch normalisation operations and red blocks represent max pooling operations. Purple blocks show fully connected layers. Each convolutional layer has a stride of (1,1), a padding size of (1,1) with a kernel size of $(3, 3)$ and is followed by ReLU activation units (not shown). Inputs are 224×224 single channel patches, as shown in Figure 2.1.

2.3.2 Network Architecture and Training

Instead of directly detecting bounding boxes we predict the corners of vertebral bodies, specifically the 4 corners of the central vertebral body in the patch and the 4 corners of the vertebral body above. To predict the corners, a variant on the VGG-F architecture [Chatfield et al. 2014] is used. However, unlike the initial formulation, batch normalisation is used between layers and each convolutional layer uses a kernel size of (3,3). A diagram of the architecture used is shown in Figure 2.3. The 16 outputs are the x and y coordinates of the 8 vertebral corners; four corners for the central vertebral body and four corners for the vertebral body above it. We choose to regress exact co-ordinates of vertebral body corners rather than the normal approach in object detection using deformable part models [Felzenszwald et al. 2008] of regressing heatmaps to the corner landmarks [Jain et al. 2014; Tompson et al. 2014; Pfister et al. 2015; Payer et al. 2016]. This is done so that the network can predict corners not visible in the patch, which is not possible with the heatmap approach.

The network is trained using mid-sagittal slices from the lumbar scans in the Genodisc training set. Each training example consists of a patch centred on a bounding box for a vertebra from S1 to L1. The coordinates for the bounding boxes of the central vertebra and the one above it are provided as labels. This patch is extracted by drawing a tightly fitting rectangle around the bounding box and then extending it by 75% on each side. Cubic interpolation is then used to resize the patch to be 224×224 pixels. This ensures constant input tensor size. Ground truth labels are extracted by applying the same transformations to the ground truth bounding box co-ordinates in the original image. The network is tasked with predicting these transformed co-ordinates using the patch as input. It is trained by minimising the L2 loss between the output vector and a 16-dimensional vector consisting of the x and y coordinates of the ground truth corners for both vertebral bodies.

The network is trained using an Adam optimiser [Kingma and Ba 2015] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 10^{-4} until convergence. Training augmentation is applied by translating the patch used by ± 10 pixels in the original image, expanding the rectangle on each side by a random amount sampled uniformly from 60% to 90% and by flipping the patches in the coronal (vertical) plane. Vertebrae decrease in size as we move up the spine so a Gaussian blur with covariance randomly selected from 3 to 30 pixels is applied to the extracted 224×224 patch. This mimics blur from resizing the relatively smaller cervical vertebral bodies at the top of the spine.

2.4 Performance Evaluation & Results

2.4.1 Evaluation

The algorithm is assessed on both lumbar and whole spine scans from the testing set of Genodisc and OWS respectively. In both cases, the algorithm was initialised by extracting the S1 vertebra in an entirely automated manner using a method derived from [Lootus et al. 2014b] and applied for 6 iterations in lumbar scans and 23 for whole spine scans. Following [Forsberg et al. 2017], a true positive detection is defined as when the centroid of a predicted bounding box is inside a ground

truth vertebral body, a false positive detection as when the centroid is not inside a vertebral body and a false negative detection as when a ground truth bounding box contains no predicted centroids.

To demonstrate the ability of the algorithm to regress bounding box corner coordinates accurately we report Dice coefficients for the overlap between predicted bounding boxes and ground truth boxes in correct detection cases.



Figure 2.4: Example vertebral extractions from mid-sagittal slices of whole spine scans in 6 different patients.

2.4.2 Results

The recall and precision of the classifier are shown in Table 2.1 for both the whole spine and lumbar spine cases. Figure 2.4 shows the examples of the algorithm applied to full spine scans.

Scan Type	Recall(%)	Precision(%)	Dice	LE (mm)
Whole Spine (All)	99.4 [1399/1408]	99.4 [1399/1408]	91.8	0.98 ± 0.68
Whole Spine (T1)	99.2 [982/990]	99.2 [982/990]	91.6	1.00 ± 0.80
Whole Spine (T2)	99.8 [417/418]	99.8 [417/418]	91.9	0.93 ± 0.61
Lumbar Spine (All)	99.8 [2872/2878]	99.8 [2871/2878]	93.5	0.62 ± 0.85
Lumbar Spine (T1)	99.6 [1344/1349]	99.6 [1344/1349]	93.2	0.65 ± 0.99
Lumbar Spine (T2)	99.9 [1528/1529]	99.9 [1528/1529]	93.8	0.58 ± 0.71

Table 2.1: Performance of the classifier on both whole spine and lumbar scans from OWS and Genodisc's test sets respectively. Results are reported on a vertebral level. The final column shows the centroid localisation error (LE).

2.5 Discussion & Conclusion

In this paper we have presented the Ladder Algorithm, a method of extracting repetitive structures from images with little training data. We have demonstrated this algorithm by using it to localise vertebral bodies in whole spine images with only annotated lumbar scans as training data.

This algorithm compares favourably to state of the art approaches at detecting vertebral bodies in lumbar scans, achieving detection rates 99.6-99.9%. For comparison, the current highest accuracy reported at this task is that of Forsberg *et al.* [Forsberg et al. 2017], which achieves 99.3-99.6% accuracy. We also report a lower localistation error of vertebrae centroids. However, it should be noted that these evaluations are done on different datasets and thus do not reflect a true comparison.

Encouragingly, the performance of this algorithm only decreases slightly when applied to whole spine scans, achieving 99.2-99.8% sensitivity and specificity. Failure cases in this example are generally due to partially visible vertebrae caused by the scanner being misaligned with the spine's orientation.

There are two major drawbacks to this approach; the algorithm runs for a fixed number of iterations meaning it is not adaptive to variations in the number of vertebrae in the spine and it relies on initialising the location of the S1 vertebra using another algorithm. Rectifying these issues is left as future work although we suggest that one solution could be to use this system to rapidly label a large amount of whole spine data and use this for training a new system robust to variable numbers of vertebrae.

Future work on this will include demonstrating the algorithm generalises to different modalities, for example CT and PET scans as well as to different tasks

including teeth and rib detection.

2.6 Acknowledgements

The authors would like to thank Timor Kadir and Andrew Zisserman for useful discussion and feedback during the researching of this paper. Rhydian Windsor is supported by Cancer Research UK as part of the EPSRC CDT in Autonomous Intelligent Machines and Systems (EP/L015897/1). Amir Jamaludin is supported by EPSRC Programme Grant Seebibyte (EP/M013774/1). The data was obtained during the ECFP7 project GENODISC (HEALTH-F2-2008-201626).

Chapter 3

A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI

This chapter presents a paper was published at MICCAI 2020, the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (Lima/Virtual).

A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI

Rhydian Windsor[1] Amir Jamaludin[1]

Timor Kadir[1,2] Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Enginnering Science, University of Oxford

[2]Plexalis

Abstract

We propose a novel convolutional method for the detection and identification of vertebrae in whole spine MRIs. This involves using a learnt vector field to group detected vertebrae corners together into individual vertebral bodies and convolutional image-to-image translation followed by beam search to label vertebral levels in a self-consistent manner. The method can be applied without modification to lumbar, cervical and thoracic-only scans across a range of different MR sequences. The resulting system achieves 98.1% detection rate and 96.5% identification rate on a challenging clinical dataset of whole spine scans and matches or exceeds the performance of previous systems of detecting and labelling vertebrae in lumbar-only scans. Finally, we demonstrate the clinical applicability of this method, using it for automated scoliosis detection in both lumbar and whole spine MR scans.

3.1 Introduction

The objective of this paper is automated vertebrae detection and identification of vertebral levels. This is an important task for several reasons. Firstly, automated diagnosis of many spinal diseases such as disc degeneration [Jamaludin et al. 2017b; Jamaludin et al. 2017d] or spinal stenosis [Lu et al. 2018] relies on accurate localisation of vertebral structures or, in the case of pathological scoliosis,

lordosis and kyphosis [Jamaludin et al. 2019], analysing the geometry of the spinal column. Secondly, vertebral bodies can be used to infer other spinal structures of interest such as the spinal cord or ribs. Finally, vertebrae can act as points to allow registration between different scans.

There are several issues that make this task challenging. One of the most obvious is that vertebrae are highly repetitive and hence distinguishing between different levels can be hard. Labelling by simply counting down from the C2 vertebra is problematic as it assumes that all vertebrae have been detected, C2 is visible, and that every patient has the same number of vertebrae which is not always true. Furthermore, for clinical use, labelling must be robust to: variations in spinal anatomy (such as collapsed vertebrae, hemivertebrae and fused vertebra); vertebrae numbers – around 11.3% of the population have one more or one less mobile vertebra [Tins and Balain 2016]); different imaging parameters including MR weighting (e.g. T1, T2, STIR, TIRM and FLAIR); fields of view (e.g. lumbar, whole spine scans); scan resolution and also number/thickness of slices in the scan.

This paper proposes a new approach to this challenge, in particular in the case of 3-D sagittal *whole spine* clinical MRIs which are important for diagnosing several diseases such as ankylosing spondylitis and multiple myeloma. We make the following contributions to the tasks of *vertebral body detection* and *vertebral level labelling* in clinical MRIs: We propose a new convolutional method of detection based on localising the corners and centroids of vertebrae and then grouping them together (Section 3.2); We reformulate the labelling task as a convolution enhancement followed by a language modelling inspired sequential correction, removing the need for a recurrent network and showing robustness to variations in vertebra numbers (Section 3.3); We show that the resulting system is robust to a variety of fields of view and pathologies by evaluating on a large clinical dataset of MR lumbar scans, and also, for the first time, on whole spine scans. We achieve state-of-the-art performance at vertebra identification in both (Section 3.4); We demonstrate a clinical application of this system by using it to automatically detect cases of scoliosis (Section 3.5).

Related work: There have been several approaches to automated detection and labelling of vertebrae in MRIs although most focus on fixed fields of view, e.g. lumbar or cervical scans only [Forsberg et al. 2017; Jamaludin et al. 2017b; Lootus

et al. 2014a; Lu et al. 2018]. Previous vertebrae labelling methods tend to rely on either heuristic-based graphical models [Forsberg et al. 2017; Lootus et al. 2013] or assuming the bottom vertebra is S1 and counting up [Lu et al. 2018]. Zhao et al. [Zhao et al. 2022] perform labelling in MRIs with arbitrary fields of view, but only in the lower spine (from S1-T12 to L4-T10). Windsor and Jamaludin [Windsor and Jamaludin 2020] report a method of detecting, though not labelling, vertebrae in full spine scans iteratively but also require the location of the S1 vertebra for initialization. Cai et al. [Cai et al. 2015] also perform detection and labelling by a hierarchical deformation model and even report success in a single full spine scan but do not evaluate the performance quantitatively on a whole spine dataset. Furthermore such models are slow to apply and make strong assumptions on spinal geometry such as a fixed number of vertebrae and lack of major pathology (fused/-collapsed vertebrae). Greater progress has been made in whole-body CT images where the task is more straightforward; CT imaging protocols are highly standardised due to image intensities being consistent, representing X-Ray absorptions of voxel locations. Also, CT scans tend to have more 3-D information, with higher resolution in the sagittal plane than that typical of clinical MR. Approaches using graphical models and recurrent neural networks for labelling have been reported to achieve 70-89% identification rate in mixed field of view CT scans [Glocker et al. 2012; Glocker et al. 2013; D. Yang et al. 2017].

3.2 Detecting Vertebrae

Detection of the vertebrae proceeds in three stages, as illustrated in Figure 3.1. First, corners and centroids of each vertebra are predicted in each sagittal slice. Second, each detected corner is assigned to a centroid by predicting a vector field for each corner type (e.g. top left, bottom right) which points to the corresponding centroid. The detected centroid and the four corners which point to it define the bounding quadrilateral for that vertebra in that slice. Third, these quadrilaterals are grouped across slices to define detected volumes for each vertebra.

Inference is performed here by a U-Net architecture (shown in detail in the appendix) which ingests an MRI (one slice) and outputs 13 channels: 5 channels are the *landmark detection* channels and each is used to detect the centroid and

4 corners respectively of all vertebral bodies appearing in that image; the remaining 8 channels are the *landmark grouping* channels corresponding to the x and y coordinates of the 4 vector fields used in grouping.

In more detail, the locations of corners and centroids are identified as modes of the heatmaps in the *landmark detection* channels. For each corner detected, there exists a corresponding grouping vector field from the *landmark grouping* channels; 4 corners equals 4 vectors fields. A group is then formed from the 4 closest corners pointing to a single centroid; a group here forms a quadrilateral. This is done for each detected centroid. If two or more centroids are assigned the same corner, the centroid to which the corner's vector points closest is retained. The other centroid is discarded to stop double detection of a single vertebra. If there is no detected corner within a fixed range of a centroid, it is also discarded, making the system robust to spurious centroid detections. This is performed in each sagittal slice. Finally, the vertebral bodies detected are grouped across slices by measuring the IoU between quadrilaterals and assigning them to the same vertebra if they have an overlap of greater than 0.5.

Discussion. Our approach of detecting vertebrae as a series of points and then grouping them differs from other methods which have used region proposal networks [Zhao et al. 2022] pixel/voxel-wise segmentation [Zukić et al. 2014], deformable-part models [Cai et al. 2015] or simply detected vertebrae centroids [Forsberg et al. 2017; D. Yang et al. 2017]. However, this approach is receiving increasing attention in the computer vision literature, with Zhou et al. [Zhou et al. 2019] showing state-of-the-art results by detecting objects as a series of keypoints. The advantage of this method is its high speed (1 second inference on GPU-enabled hardware) combined with more accurate bounding regions than using standard bounding boxes.

Training: To train the network, we use separate loss functions for the detection and grouping channels. The ground truth annotations are coordinates of the corners of the vertebral bodies (centroids are computed from these). Target detection channel outputs are constructed by overlaying a Gaussian kernel on each annotated ground truth landmark in the detection channels with variance proportional to the square root of the area of the detection. These channels are trained using the weighted L1 loss; $\mathcal{L}_{detect}(Y, \hat{Y}) = \sum_{k=1}^{5} \sum_{i,j} \alpha_{ijk} |y_{ijk} - \hat{y}_{ijk}|$ where Y is the response

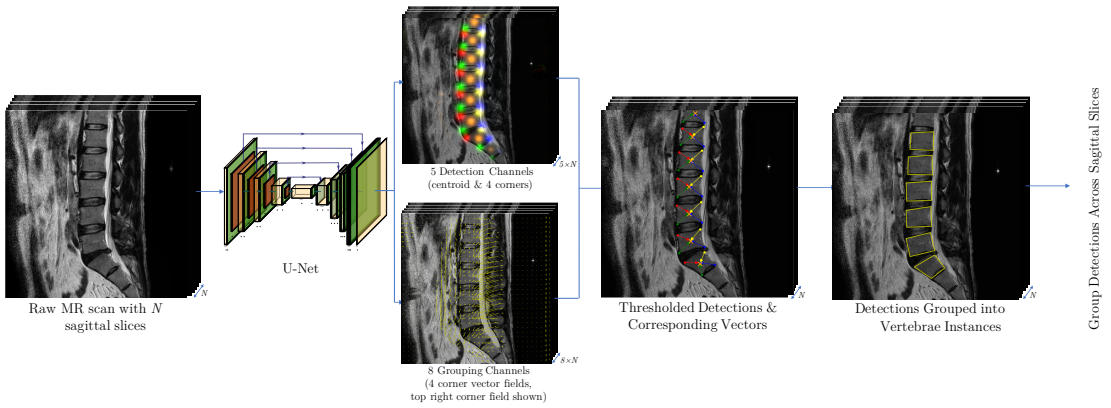


Figure 3.1: The pipeline used to detect vertebrae in whole spine scans. The output detection channels are thresholded and each resulting connected volume becomes a corner or a centroid. Each corner has corresponding vector in the grouping channels at the point-of-detection and 'points' with a magnitude and direction according to that vector. Each centroid is assigned the 4 corners that point closest to it.

map output by the network and $\hat{Y}$ is the target response map. y_{ijk} and $\hat{y}_{ijk}$ are the value of response and ground truth map respectively at image coordinate (i, j) in detection channel k and a_{ijk} is a weighing factor given by $\alpha_{ijk} = \frac{N_k}{N_k+P_k}$ if $\hat{y}_{ijk} \geq T$ or $\frac{P_k}{N_k+P_k}$ if $\hat{y}_{ijk} < T$, where P_k and N_k are the number of pixels respectively above and below threshold T in channel k. This weighing factor balances the loss from false positive responses in the heatmap and false negatives, speeding up training. In the experiments in this paper, $T = 0.01$ was used. *Landmark grouping* channels are trained by using the L2 loss; $\mathcal{L}_{group} = \sum_{l=1}^{4} \sum_b \sum_{(i,j)\in\mathcal{N}_{bl}} ||\mathbf{v}_{ij}^l - \mathbf{r}_{ij}^k||_2^2$. Here l indexes each corner type/vector field (e.g. top left, bottom right), b indexes labelled ground truth vertebral body, $\mathcal{N}_{bl}$ is a neighbourhood of pixels surrounding the l^{th} corner of vertebral body b. $\mathbf{v}_{ij}^l$ is the value of the vector field corresponding to corner l at the pixel location (i, j) and $\mathbf{r}_{ij}^b$ is the displacement from the centroid of vertebral body b to (i, j). As suggested by [Ronneberger et al. 2015], heavy augmentation is used during training. Scans are padded, rotated, zoomed and flipped in the coronal plane. Non-square scans are split into squares overlapping by 40% and resized to 224×224 to ensure constant sized input to the network.

3.3 Labelling Vertebrae

Once the vertebrae are detected, the next task is to label each vertebrae with its level (e.g. S1, L5, L4 etc.). There are two types of information to consider when

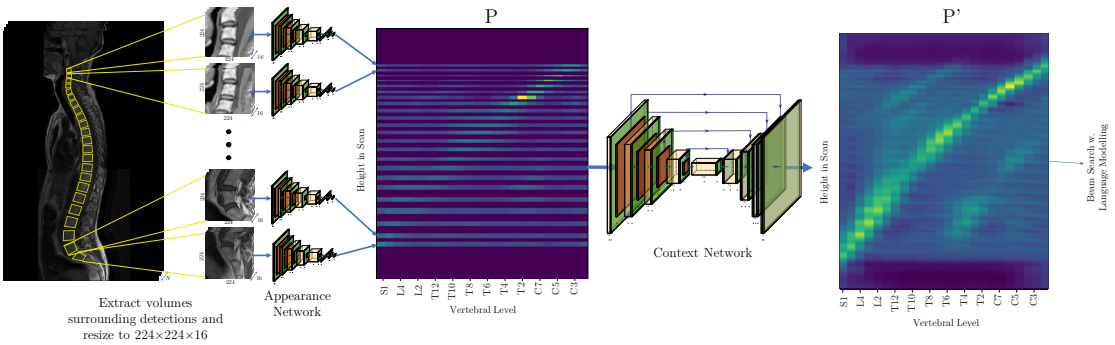


Figure 3.2: Overview the labelling pipeline. Volumes surrounding each of the vertebrae detected are extracted and input to an appearance network. This is then used to construct an input image to the context network which gives the probability of a given label at a given height in the image. The resulting probability map is shown in the far right of the image, followed by a beam search to generate the final level sequence. The architectures of both networks are shown in the appendix.

labelling a vertebra; the *appearance* of the vertebrae – its intensity pattern, shape and size – and its *context* – its position in relation to other vertebrae that have been detected in the scan. As such, we train two networks; an *appearance network* to infer the level of vertebra from its appearance alone, and a *context network* which takes as input the predictions of the appearance network along with the spatial configuration of the detections in the scan to improve the predictions. The final stage is to search for a consistent labelling of the vertebra using a sequential 'language model' that builds in ordering constraints. The labelling pipeline is outlined in Figure 3.2. It should be noted that both networks are fully convolutional. This differs from the approaches outlined in [D. Yang et al. 2017] and [Lootus et al. 2014a] which use recurrent neural networks and graphical models respectively.

The labelling pipeline proceeds as follows: A volume around each detected vertebra is extracted and given as input to the appearance model which predicts a softmax probability vector over the labels. Input volumes are created by first fitting a bounding box around the detection, then expanding the box by 100% to include nearby anatomical features and resampling its size to $224 \times 224 \times 16$, where 16 is the number of sagittal slices. For the appearance network, we use a simple 3D CNN (details of this and the context network are given the appendix). Then, the predictions of the appearance networks are used to make a probability-height map where the probability vector at the height of each vertebra is equal to the output vector of the appearance network for that vertebra. In detail: the probability-height map, $P \in \mathbb{R}^{H \times N}$ where H is the height of the image and $N = 24$ is the

number of vertebrae level classes (C2 to S1). If vertebra detection v spans from heights h_{v_1} to h_{v_2} in the scan then $P_{h'n} = p^a_{vn} \forall\ h_{v_1} \leq h' \leq h_{v_2}$ and 0 otherwise where p^a_{vn} is the probability v has level n, given by the appearance network. Next, the context network refines this probability-height map, using detections around each vertebra to update its predicted class. Finally, the output map is decoded into a logical sequence of vertebra using a language-model inspired beam search which imposes global constraints such as no repetitions. This is described in detail in Section 3.3.1.

Training: The appearance network is trained using a cross-entropy loss function. For re-calibration [Guo et al. 2017], a softmax operation with temperature $T = 10$ is applied to the logits layer. P is given to the context net which produces P', a refined version of the same map as shown on the far right of Figure 3.2. The cross-entropy between the output probability map at the centroid height of each vertebra and its ground truth label is used as the loss function. When training the context network, augmentation is applied at training time by randomly removing between 0 to 4 of the highest and lowest detections from the input image and further removing each remaining vertebra's probability map from the input image with probability 0.2. This ensures the labelling pipeline is robust to missed detections and not reliant on distinct looking vertebrae at the top and bottom of the spine.

3.3.1 Enforcing Monotonicity and Dealing with Numerical Variations

To predict the labels for each vertebrae detected, P' shown in Figure 3.2 must be decoded into a sequence of level predictions. A naive method to do this would be to take the argmax of the probability map at each vertebra centroid height and use this as a label. However, this allows implausible sequences such as repeated vertebral levels to be predicted. If the detection v at height h_v has label index l, then we know v' at $h_{v'} > h_v$ detection should have label $l' > l$ with the greatest probability of being $l + 1$(e.g. L4 should be followed by L3). The challenge of imposing such constraints is analogous to that faced in automatic speech recognition where CTC training is followed by language modelling to predict a valid character or word sequence [Scheidl et al. 2018; Graves et al. 2006]. We take inspiration from

this and use a beam search to obtain a valid labelling: beginning at the highest detection and searching down, sequences are generated and scored by selecting the highest probability labels for each vertebra. The k most likely sequences of levels are stored in memory at each step, where k is the beam width. Sequences with repetitions of levels are given probability 0 and those with skipped levels are penalised by multiplying the sequence probability by a penalty score. The method can also incorporate numerical variations, with ± 1 lumbar vertebrae given a small sequence probability penalty reflecting the low incidence of this.

3.4 Detection and Labelling Results

We evaluate the system at the task of vertebrae detection and labelling in whole spine and lumbar scans. Following [Lootus et al. 2013; D. Yang et al. 2017], we define correct detections to be when the ground truth vertebra centroid is contained entirely within a single bounding quadrilateral. For detection, we report precision, recall, and the localisation error (LE), defined to be the mean distance of ground truth centroids from the closest detection quadrilateral centroid. For labelling, we report identification rate (IDR), the fraction of vertebrae detected and labelled correctly.

3.4.1 Datasets

Three datasets are used in this work: OWS, Genodisc and Zukić. OWS is a dataset of 710 sagittal whole spine scans across 196 patients from the Picture Archiving and Communication System (PACS) of an orthopaedic centre. The dataset exhibits a wide range of pathologies such as hemivertebra, fused vertebrae, numerical variations of vertebrae and scoliosis. Scans are taken from different scanners with a range of MR parameters (T1, T2, FLAIR, TIRM and STIR). The dataset is split into training, validation and testing sets with an 60/20/20% split at the patient level. Corners of vertebrae from S1 to C2 are annotated and vertebrae levels were marked by a radiologist in one scan for each patient, with S1 being the first vertebra attached to the pelvis. In the case of 25 vertebrae between S1 and C2 instead of the normal 24, an extra lumbar vertebrae is labelled (L6). Networks

Dataset	# Scans	# Vert	Method	Precision (%)	Recall (%)
OWS (Whole Spine)	37	888	[Windsor and Jamaludin 2020]†	**99.4**	**99.4**
			Label Baseline	-	-
			Ours	99.0	98.1
Genodisc (Lumbar)	421	2947	[Lootus et al. 2013]	-	-
			Label Baseline	-	-
			Ours	**99.7**	**99.7**
Zukić (Lumbar)	17	154	[Zukić et al. 2014]	98.7	92.9
			Label Baseline	-	-
			Ours	**99.3**	**98.7**

Dataset	# Scans	# Vert	Method	IDR(%)	IDR± 1(%)	LE (mm)
OWS (Whole Spine)	37	888	[Windsor and Jamaludin 2020]†	-	-	**1.0 ± 0.9**
			Label Baseline	86.9	93.4	-
			Ours	**96.5**	**97.3**	2.4 ± 1.3
Genodisc (Lumbar)	421	2947	[Lootus et al. 2013]	86.9	-	3.5 ± 3.3
			Label Baseline	90.1	97.4	-
			Ours	**98.4**	**99.7**	**1.6 ± 1.1**
Zukić (Lumbar)	17	154	[Zukić et al. 2014]	-	-	**1.6 ± 0.8**
			Label Baseline	87.0	94.3	-
			Ours	**90.9**	**98.7**	2.0 ± 1.5

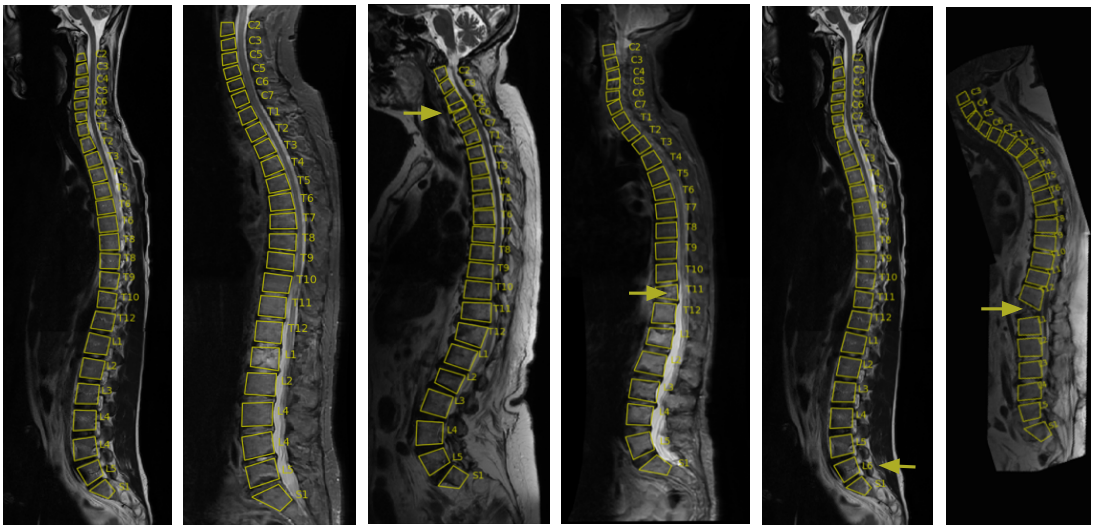
Table 3.1: Performance of the pipeline on the three datasets. Our approach is compared with other methods using the same datasets and also a LSTM labelling baseline. Results are reported on a per-vertebra level. Higher is better for detection precision, detection recall, and level identification rate (IDR). Lower is better for localisation error (LE). We also report the percentage of vertebrae within one level of their ground truth value (IDR±1).[Lootus et al. 2013] is tested on a subset of 291 scans from the Genodisc dataset. **Note, [Windsor and Jamaludin 2020]† requires manual initialization by providing the location of the S1 vertebra (either by human selection or an S1-specific detection system), so is not directly comparable.**

were trained on the OWS training set, using an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.001. The Genodisc and Zukić datasets are used only for testing. Genodisc's test set has 421 clinical lumbar MRIs used by Lootus et al. [Lootus et al. 2014a]. Zukić [Zukić et al. 2014] is a small dataset of 17 mostly lumbar sagittal MRIs available on the online SpineWeb platform.

3.4.2 Results

The results of detection and labelling on all datasets are shown in Table 3.1 with comparisons to other methods reported on the same datasets where available. We also compare our convolutional labelling pipeline results to a baseline recurrent approach for vertebra labelling, training a bidirectional LSTM on the appearance features extracted from each detected volume. Example predicted detection and labelling sequences across a range of pathologies are given in Figure 3.3. The LSTM baseline used is detailed in the appendix.

Discussion: For whole spine detection on the OWS dataset, the proposed method



a) b) c) d) e) f)

Figure 3.3: Example detection and labelling of vertebrae for a range of whole spine scans. a) and b) are examples of typical spines. The other images shows examples of the system dealing with pathologies; c) shows an example of a spine with fused C3 and C4 vertebrae; d) is an example of a spine with a collapsed vertebra; e) shows a spine with an extra vertebra between S1 and C2 and f) shows an extremely kyphotic spine with a hemivertebra at the bottom of the thoracic spine.

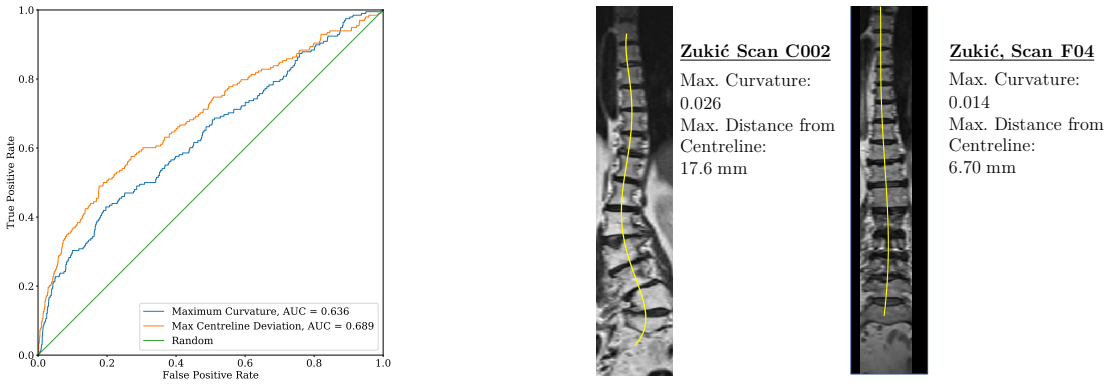
achieves a high precision and recall of 99.0% and 98.1% respectively. It achieves a level identification rate of 96.5%, significantly exceeding the LSTM baseline. The few errors the system made for labelling are generally due to S2 being detected as S1, meaning all labels are out by one. In practice, this is a mistake radiologists often make as it can be difficult to tell which is the first sacral vertebra without looking at axial scans to see which bone is joint to the pelvis. This also explains why [Windsor and Jamaludin 2020] achieves slightly higher precision and recall at detection than ours; it is a semi-automated algorithm given the location of S1 at initialization and thus bypasses this difficult problem of S1 recognition. On the Genodisc lumbar spine dataset, our method again significantly outperforms the baseline, and it is also outperforms the prior method of Lootus et al. [Lootus et al. 2013] (98.4% compared to 86.9%). Importantly, OWS and Genodisc are scans of patients with a wide range of pathologies imaged using typical clinical MR protocols; hence strong performance here gives evidence of the clinical usefulness of this approach. We show example results for pathological spines in the appendix. In the appendix we also report results by other groups for vertebrae detection and labelling in MRIs. However, these are for different datasets to which we could not

get access, with different scanning protocols, fields of view (FoV) and patient sets, and thus cannot be compared to directly.

3.5 Automated Scoliosis Detection

Finally, as an illustration of a potential application of the proposed approach, we explore the ability of the system to classify cases of scoliosis from sagittal MR scans. In clinical practice this is usually determined by measuring the Cobb angle in coronal views of X-ray scans [Aebi 2005], however measuring scoliosis in the supine position has also been shown to be possible [Taylor et al. 2013; Jamaludin et al. 2019] and MRI can be useful for understanding disease etiology and symptoms [Ozturk et al. 2010]. This is a more difficult task in sagittal scans as it requires sensitive detection of the sides of vertebra with a clear decision boundary as to when a vertebra is present or not present in a slice which can be difficult in cases of partial visibility. In the entirety of the Genodisc dataset, scoliosis was reported by a radiologist in 198 of 3542 scans, across 2009 patients. By measuring statistics of a quintic polynomial fit through vertebra centroids and using them as predictive features we develop classifiers for this label. Specifically, we measure the maximum curvature of the polynomial, and the maximum deviation of the curve from a straight vertical line fit through the vertebrae, assuming that a low curvature vertebral column with little deviation from the centreline corresponds to a non-scoliotic spine. While ground truth scoliosis labels were not available for the whole spine scans, we give qualitative results, comparing the features of curves fit though vertebrae of a scoliotic and non-scoliotic scan from Zukić. Results of these experiments are shown in Figure 3.4.

Discussion: The results for automated scoliosis detection from sagittal scans are promising. Using simple classifiers AUCs of 0.636-0.689 are achieved in a highly class-imbalanced problem. Of the features measured, distance from the vertical centreline of the vertebrae performed best. The system is also shown to capture scoliotic curves in full spine scans too. These experiments illustrate that the proposed method produces a strong geometric representation of the spine, which can be used in further downstream tasks.



(a) (b)

Figure 3.4: Results of scoliosis classification in sagittal scans using the proposed vertebrae detection system: (a) ROC curve for simple scoliosis classifiers based on statistics of polynomial curves fitted through vertebrae centroids in Genodisc scans; (b) A qualitative comparison of curves fit through detected vertebrae in full spine scans. Curves are overlaid on coronal slices synthesised from the sagittal slices.

3.6 Conclusion

We introduce a novel method for vertebrae detection and labelling in whole spine sagittal MRIs. It shows state-of-the-art results for vertebra identification in lumbar scans, with little performance drop in the whole spine case, and is robust to a range of spinal defects, numerical variations of vertebrae and different scanning protocols. We also demonstrate a potential diagnostic application: automated detection of scoliosis from sagittal MRIs. Future work will include integrating automated detection of different spinal pathologies and implementing the system for CT scans.

Acknowledgements. The authors would like to thank Dr. Sarim Ather for useful discussions on spinal anatomy and clinical approaches to diagnosing disease, as well as assistance labelling the data. Rhydian Windsor is supported by Cancer Research UK as part of the EPSRC CDT in Autonomous Intelligent Machines and Systems (EP/L015897/1). Amir Jamaludin is supported by EPSRC Programme Grant Seebibyte (EP/M013774/1). The Genodisc data was obtained during the EC FP7 project GENODISC (HEALTH-F2-2008-201626).

3.7 Appendix

3.7.1 Overview of Proposed Pipeline

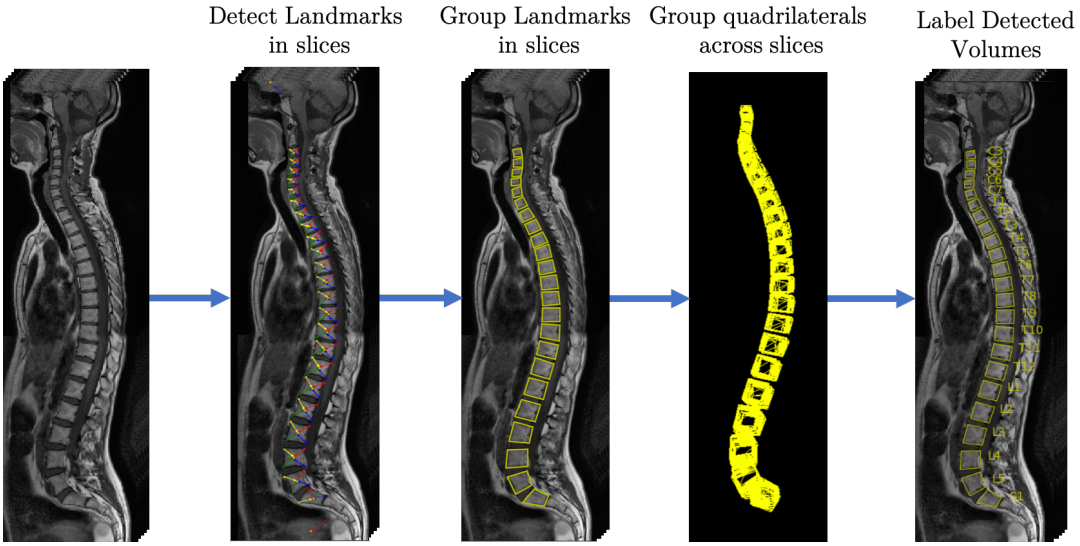


Figure 3.5: An overview of the proposed detection and labelling pipeline. Vertebrae landmarks are detected, grouped into quadrilaterals in each sagittal slice and finally grouped across slices. The output volumes are input into the labelling pipeline.

3.7.2 LSTM Baseline Architecture

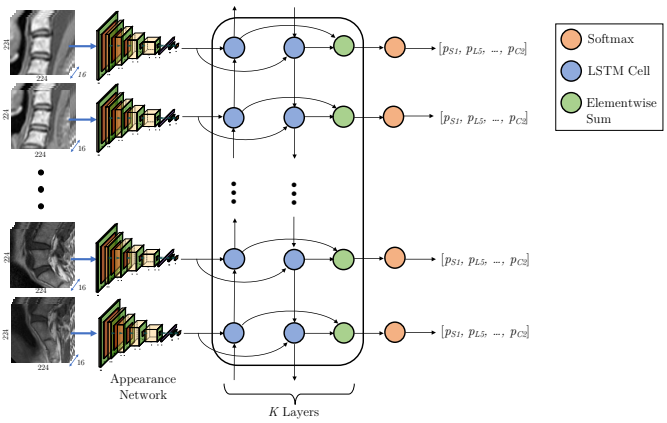


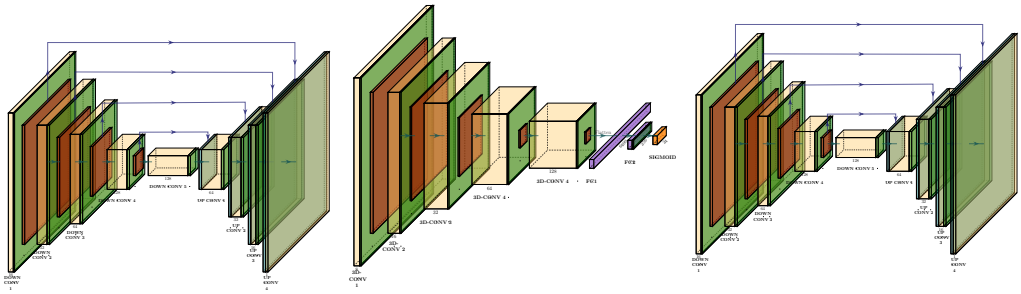
Figure 3.6: The LSTM architecture used as a labelling baseline. Output probability maps from the appearance network are used as input to a bidirectional LSTM, with the inputs ordered by the height of the vertebrae in the image. The architecture outputs assigns a probability vector for each detection which the likeliness of the vertebra being each level from S1 to C2. In the experiments in Section 4, $K = 2$ is used.

3.7.3 Other Reported Results

Table 3.2: Reported results of other methods proposed for the segmentation of vertebrae in CT and MR spinal scans evaluated on different datasets. We attempted to access a publically available dataset from [2] although did not get a reply prior to submission.

Method	FoV		Comment	
Cai et al. [2]	Arbitrary		MR & CT, Deformable shape model	
Forsberg et al. [4]	Lumbar, Cervical		MR, requires S1 or C2 visibility	
Glocker et al. [4]	Mostly Abdominal		CT, Random Forest Classifier	
Glocker et al.[5]	Mostly Abdominal		CT, Random Forest Classifier	
Yang et al. [18]	Arbitrary		MR, LSTM labelling	
Method	Prec. (%)	Rec. (%)	LE(mm)	IDR (%)
Cai et al. [2]	-	-	2.54-3.12	92-97
Forsberg et al. [4]	99.6-100	99.1-99.8	1.18-2.60	96-97
Glocker et al. [4]	-	-	5.94-8.53	72-85
Glocker et al.[5]	-	-	7.00-14.3	62-86
Yang et al. [18]	-	-	6.90-9.00	83-89

3.7.4 Detection, Appearance and Context Network Architecture Diagrams



(a) The detection network. All layers have 3×3 kernels with 1×1 padding. MaxPool operations have a kernel size of 2×2.

(b) The appearance network used to label vertebrae volumes. The first two layers have kernel sizes of $3 \times 3 \times 1$ with $3 \times 3 \times 3$ in later layers.

(c) The context network. All Conv layers have 5×3 kernels with 2×1 dilation, 4×1 padding and 2×1 stride. MaxPool operations have 2×1 kernels.

Figure 3.7: The network architectures used in the detection and labelling pipeline.Yellow blocks indicate convolutional layers, green blocks indicate batch normalisation, purple layers are fully connected and orange layers are sigmoid activation units. All layers except the last use ReLU activation units (not shown).

3.7.5 Detection and Labelling Accuracy Across Different Vertebral Levels

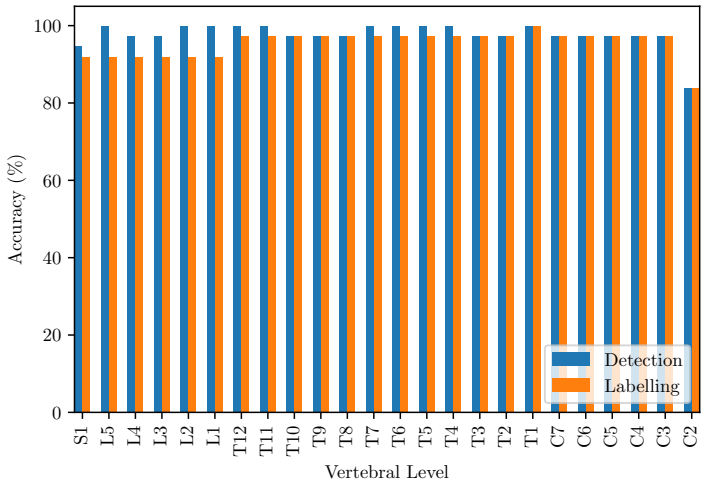


Figure 3.8: Detection recall and vertebral identification rate across different vertebral levels in the OWS whole spine dataset. The system is least accurate at detecting vertebrae at the extremities of the vertebral column (S1 and C2).

Chapter 4

Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading

This chapter presents a paper published at MICCAI 2022, the 25$^{\text{th}}$ International Conference on Medical Image Computing and Computer Assisted Intervention (Singapore).

Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading

Rhydian Windsor[1] Amir Jamaludin[1]

Timor Kadir[1,2] Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Enginnering Science, University of Oxford

[2]Plexalis

Abstract

This paper proposes a novel transformer-based model architecture for medical imaging problems involving analysis of vertebrae. It considers two applications of such models in MR images: (a) detection of spinal metastases and the related conditions of vertebral fractures and metastatic cord compression, (b) radiological grading of common degenerative changes in intervertebral discs. Our contributions are as follows: (i) We propose a Spinal Context Transformer (SCT), a deep-learning architecture suited for the analysis of repeated anatomical structures in medical imaging such as vertebral bodies (VBs). Unlike previous related methods, SCT considers all VBs as viewed in all available image modalities together, making predictions for each based on context from the rest of the spinal column and all available imaging modalities. (ii) We apply the architecture to a novel and important task – detecting spinal metastases and the related conditions of cord compression and vertebral fractures/collapse from multi-series spinal MR scans. This is done using annotations extracted from free-text radiological reports as opposed to bespoke annotation. However, the resulting model shows strong agreement with vertebral-level bespoke radiologist annotations on the test set. (iii) We also apply SCT to an existing problem – radiological grading of inter-vertebral discs (IVDs) in lumbar MR scans for common degenerative changes. We show that by considering the context of vertebral bodies in the image, SCT improves the accuracy for several

gradings compared to previously published models.

4.1 Introduction

When a radiologist is reporting on a patient's imaging study, they will typically make assessments based on multiple scans of the same patient. For example, in a MRI study they will often consider multiple pulse sequences showing the same region and note differences between them. Related conditions may appear similar in one modality and only be distinguishable when more context is given via an additional imaging modality. Similarly, when reporting on spinal images, radiologists can learn a lot about a specific vertebra based on its appearance *relative to the vertebrae neighbouring it*. For example, a region of hyperintensity may represent a lesion or may just be an artifact of the scanning protocol used; this often can be elucidated by looking at the rest of the spine.

Accordingly, automated models for spinal imaging tasks should also be able to make vertebra-level predictions with context from multiple sequences *and from neighbouring vertebrae*. In this work, we propose the *Spinal Context Transformer* (SCT), a model which aims to do exactly this. Crucially, SCT leverages lightweight transformer-based models which allow for variable numbers of inputs (both during training and at test-time) in terms of both: (a) the number of vertebral levels, and (b) the number of input modalities depicting each vertebra.

As well as introducing SCT, this paper also explores a new application of deep learning in medical imaging: automated diagnosis of spinal metastases and related conditions. This is an important task for several reasons. Firstly, metastases are very common; overall 2 in 3 cancer cases will metastasise [Shaw et al. 1989] (rising to nearly 100% incidence in patients who die of cancer [Maccauro et al. 2011]). The spine is one of the most common places for these metastases to occur [Ortiz Gomez 1995]. Secondly, early detection is vital; metastases generally indicate advanced cancer and, if left untreated, cause conditions such as vertebral fractures and metastatic cord compression, both of which can lead to significant pain and disability for the patient and are difficult and expensive to treat [van Tol et al. 2021a; van Tol et al. 2021b]. Current clinical practice relies on heuristic-based

scoring systems to quantify the development of metastases, such as SINS [Fisher et al. 2010] and the Tokuhashi score [Tokuhashi et al. 2014]. We aim to develop models to aid in the rapid detection and consistent quantification of this serious condition. To further validate our model, we also test it on a previously published dataset for grading common intervertebral disc (IVD) degenerative changes in lumbar MR scans.

Another theme explored in this work is the ability of free-text radiological reports to generate supervisory signals for medical imaging tasks. The common praxis when training models on radiological images is to extract a relevant dataset from an imaging centre, pseudo-anonymise it and then annotate the dataset for a condition of interest. While this method is simple, effective and very popular, the additional step of annotating the dataset can be a significant bottleneck in the size of datasets used. This is a more serious problem than in conventional computer vision datasets as medical image annotation generally requires a specialist, whose time is limited and expensive. A much more scalable method of curating annotated datasets is to use existing hospital records for generating supervisory signals. However, in radiology, these records are usually in the form of free-text reports. Due to variation and lack of structure in reporting, extracting useful information from these reports has been considered a difficult task [Casey et al. 2021], and, as such, most studies annotate their own data retrospectively. In this paper, we explore how useful the information in free-text reports can be for training models for our specific task of detecting spinal cancer, and propose methods for dealing with ambiguities and omissions from the text.

4.2 Related Work

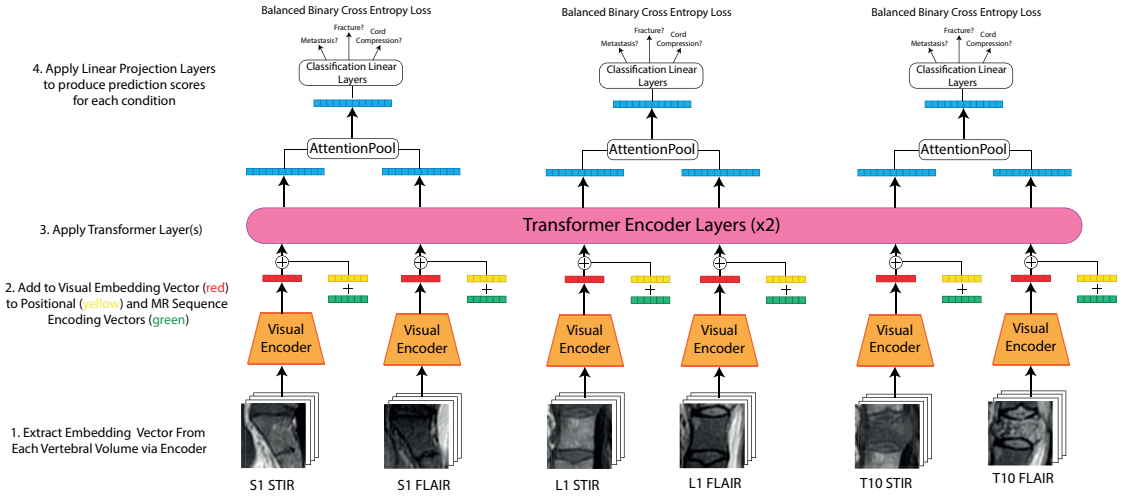
The SCT, and its training and application, builds on three related areas of work: deep learning for analysis of spinal scans, obtaining supervision from free-text radiology reports and transformers for aggregating multiple sources of information. Automated detection and labelling of vertebrae is a long standing task in spinal MRI [Lootus et al. 2013] and CT [Glocker et al. 2012] analysis, with deep learning now the standard approach [Cai et al. 2015; Forsberg et al. 2017; Tao and Zheng 2021; Windsor et al. 2020; Windsor and Jamaludin 2020]. One study particularly

relevant to this work is that of Tao *et al.* [Tao and Zheng 2021] who use a transformer architecture for this task. Beyond vertebra detection and labelling, deep learning has also been used to assess spinal MRIs for degenerative changes. For example, SpineNet of Jamaludin *et al.* [Jamaludin et al. 2017b] and Windsor *et al.* [Windsor et al. 2022] is a multi-task classification model acting on intervertebral discs which grades for multiple common conditions. We compare to this approach here. DeepSpine of Lu *et al.* [Lu et al. 2018] and also Lewandrowski *et al.* [Lewandrowskl et al. 2022] use ground truth derived from radiological reports to train models for grading stenosis and disc herniation respectively, a theme we also explore in this work. In terms of spinal cancer, [Zhao et al. 2022] and [J. Wang et al. 2017] both train models to detect metastases based on 2D images extracted from MRI studies; and [Merali et al. 2021] proposes a model to automatically detect metastatic cord compression, although this is restricted to axial T2 scans or the cervical region. Several other works have considered vertebra and metastases detection in other modalities such as CT (e.g. [Hammon et al. 2013; Burns et al. 2013; Glocker et al. 2012; Glocker et al. 2013]), although MR remains the clinical gold-standard in early spinal cancer detection. Finally, our approach also has parallels with works in video sequence analysis that combine a 2D CNN backbone with a temporal transformer for tasks such as representation learning, tracking, object detection and segmentation [Bain et al. 2021; Gabeur et al. 2020; Carion et al. 2020; Meinhardt et al. 2022]; though in our case, visual features are extracted from multiple vertebrae shown in multiple MR sequences rather than consecutive frames.

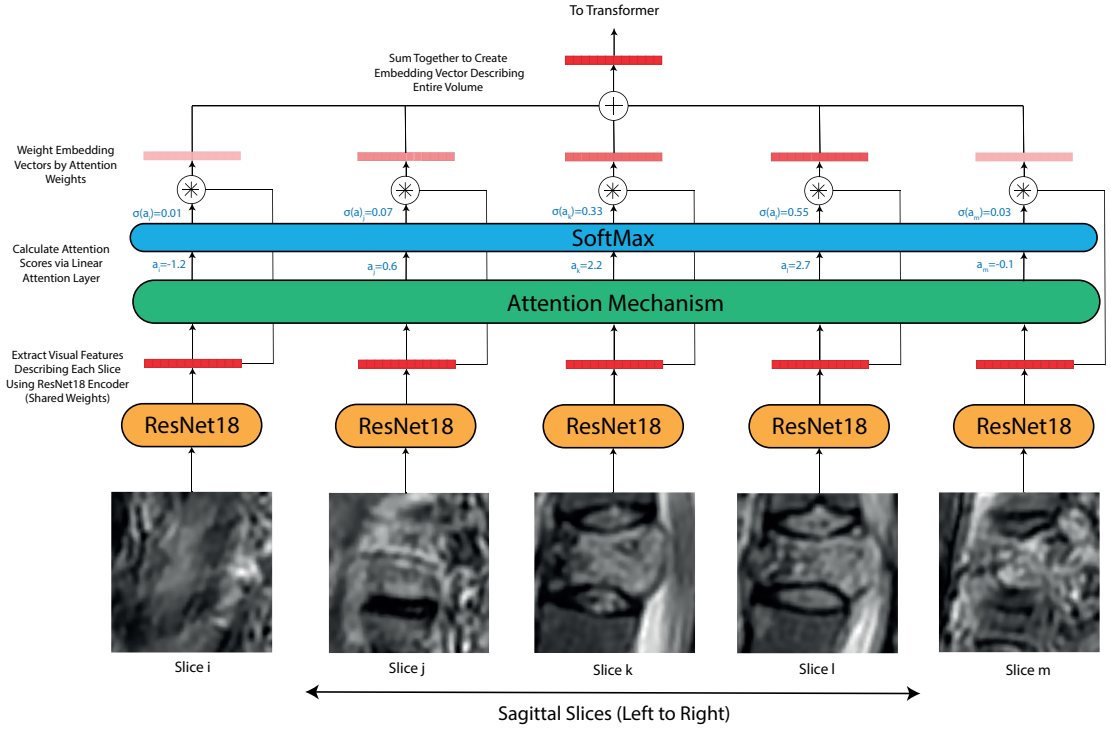
4.3 Spinal Context Transformer

Like other models used to assess vertebral disorders such as SpineNet [Jamaludin et al. 2017b; Windsor et al. 2022] and DeepSpine [Lu et al. 2018], the Spinal Context Transformer (SCT) exploits the repetitive nature of the spine by using the same CNN model to extract visual features from each vertebra given as input. However, there are two key conceptual differences between how SCT and these other models operate which are described in this section.

Firstly, instead of using 3D convolutions and pooling layers to extract features from



(a) The SCT architecture illustrated for the spine cancer prediction task operating on multiple MRI sequences of three vertebrae. In the actual cancer detection task, the whole spinal column is used, and SCT is able to handle arbitrary variations in the number and type of input sequences and vertebrae without adaptation.



(b) The visual encoder network used to extract features from each vertebral volume, This model can be trained by itself or in conjunction with the transformer architecture as shown in (a). The same model weights are used for all MR sequences.

Figure 4.1: Architecture diagrams for SCT: (a) the full SCT architecture; (b) the encoder used to extract visual features from each vertebra. STIR and FLAIR sequences are shown here, however the radiological grading task uses strictly T1 and T2 sequences, whereas the spinal cancer task uses a range of common sequences (T1, T2, STIR, FLAIR etc).

multislice MR images, SCT instead extracts features from each slice independently using a standard 2D ResNet18 [He et al. 2016] architecture adapted for single-channel images. The extracted feature vectors for each slice are then aggregated

using an attention mechanism. This is shown in Figure 4.1b. This design choice is made for several reasons: (i) 3D convolutions assume all images have the same slice thickness. In real-world clinical practice a variety of scanning protocols are used. This means slice thickness can vary significantly from sample to sample. Of course, resampling can be employed in pre-processing to ensure a consistent slice thickness for input images, however this introduces visual artifacts which may degrade performance. (ii) Trained encoder models can operate on 2D or 3D scans without adaptation. This also means image CNNs trained on large 2D datasets such as ImageNet [Deng et al. 2009] can be used to initialize the encoder backbone. (iii) The attention mechanism forces the model to explicitly determine the importance of each slice for its prediction. This allows peripheral slices with partial-volume effects to be ignored and can be useful when it comes to interpreting the output of the model. An example of this effect is shown in the appendix. Secondly, as well as using attention to aggregate feature vectors from each slice, SCT also uses attention to collate information from each MR sequence and vertebra in the spinal column. This is done by feeding the vertebra's visual embedding vectors from each MR sequence to a lightweight 2-layer transformer encoding model, with additional embedding vectors describing the level of each input vertebra and the imaging modality used. Both these additional embeddings are calculated by linear layers operating on a one-hot encoding of the vertebra's name and sequence. The output feature vectors for the same vertebra shown in different sequences are then pooled together using a final attention mechanism identical to that used to pool features across slices. This creates a single output vector for each vertebra. Linear classification layers convert this output vector into predictions for each classification task. The full process is shown in Figure 4.1a. Annotated PyTorch-style pseudo-code is given in the appendix.

4.4 Detecting Spine Cancer Using Radiological Reports

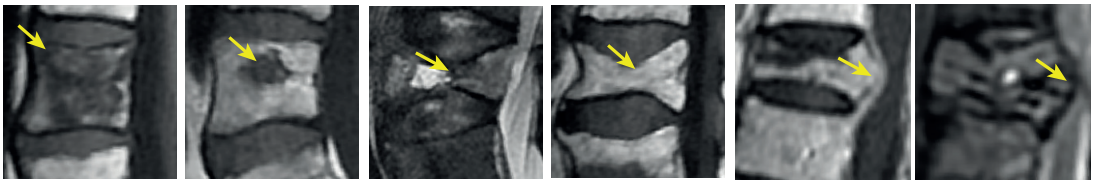
This section describes applying SCT to a novel task: detecting the presence of spinal metastases in whole spine clinical MR scans, as well as the related conditions of vertebral fractures/collapse and spinal cord compression, employing in-

formation from free-text radiological reports for supervision. We use a dataset of anonymised clinical MRI scans and associated reports extracted from a local hospital trust PACS (Oxford University Hospitals Trust) with appropriate ethical clearance. Inclusion criteria were that the patient was over 18 years old, had at least one whole spine study, and was referred from a cancer-related specialty. In order to ensure a mix of positive and negative cases, all patients that matched this criteria from April 2015 until April 2021 were included in the dataset, regardless of whether or not they had metastases.

To avoid having to annotate each scan independently we instead rely on the radiological reports written at the time the scan was taken to provide supervision for training our models. A major advantage of this approach is that it is much faster and thus readily scalable to larger datasets – since no radiologist is required to review each image independently, an annotation set can be obtained much quicker than would otherwise be possible. Furthermore, significantly less clinical expertise is required to generate these annotations – only a basic understanding of vocabulary related to the conditions of interest is needed.

A particular challenge of using annotations derived from reports for training is dealing with ambiguities in the text. Firstly, reports are often inconclusive; e.g. "The nature of this lesion is indeterminate" or "Metastases are a possibility but further investigation is required". In these cases we label vertebrae as 'unknown' for a specific condition and do not apply a loss to the model's corresponding predictions. Note if one vertebra is marked as positive in a spine and the other levels are ambiguous, then the others are marked as 'unknown'. This accounts for cases where a specific metastasis/fracture is being commented on due to a change from a previous scan but other unchanged metastases/fractures are not fully described. Cord compression, on the other hand, is marked as negative unless explicitly stated otherwise, as it is a severe condition which is highly unlikely to go unmentioned in a report. Secondly, reports often state that "metastatic disease is widespread" throughout the spine but do not indicate at which specific levels this disease occurs. To derive annotations from these ambiguous reports, we introduce an additional, global label for each study, indicating if the given condition is present in the scan at any vertebral level or not. We then aggregate each vertebra-level prediction to produce a spinal-column level prediction of whether the condition is present. This

setting is an example of multiple-instance learning (MIL), whereby instead of each training instance being individually labelled, sets of instances (known as 'bags') are assigned a single label, indicating if at least one instance in the bag is positive. The challenge is then to determine which instances in the bag are positives. In this case each bag represents an entire spinal column and each instance is a single vertebra. We can then train by a hybrid single-instance and multiple-instance learning approach; vertebra-level annotations are used for supervision where given by the report in addition to the scan-level annotations given in each case. A breakdown of the labels extracted by this method is given in Table 4.1. Examples of vertebral bodies from each class are shown in Figure 4.2. Further explanation and example annotations are given in the appendix. To ensure our model produces results similar to those given by bespoke annotations, our test dataset is labelled by an expert spinal surgeon.



(a) Metastases (b) Collapsed Vertebrae (c) Cord Compression

Figure 4.2: Example vertebrae with the three spinal-cancer related conditions of interest. Note that in many cases vertebrae have a combination of these conditions.

Split	Patients	Studies	Total	Vertebral Bodies								
				Metastases			Fractures			Compression		
				+	−	?	+	−	?	+	−	?
Training	258	558	12459	829	6099	5531	216	12124	119	90	12369	0
Validation	33	63	1391	81	829	482	25	1360	7	15	1377	0
Test (Reports)	32	66	1430	92	646	692	22	1408	0	11	1419	0
Test (Expert)	32	66	1430	374	1056	0	46	1384	0	31	1399	0

Table 4.1: The dataset used to train SCT to detect metastases, fractures and compression. The first-three rows indicate the number of labels extracted from free-text reports for the training, validation and test splits. The bottom row indicates the same test dataset with each vertebra independently annotated by an expert. Vertebra are labelled as positive (+), negative (−), or unknown (?). Note the expert annotator was required to make their best guess in cases of uncertainty.

4.5 Radiological Grading Of Spinal Degenerative Changes

To validate SCT on an existing vertebra analysis problem, we also measure its performance at grading common degenerative changes around the intervertebral discs.

Specifically, we grade the following common conditions: Pfirrmann Grading, Disc Narrowing, Endplate Defects (Upper/Lower), Marrow Changes (Upper/Lower), Spondylolisthesis and Central Canal Stenosis. We do this on Genodisc, a dataset of clinical lumbar MRIs of 2295 patients from 6 different clinical imaging sites. This problem setting closely follows that discussed in [Jamaludin et al. 2017b] and [Windsor et al. 2022], which give the existing state-of-the-art methods for many of these grading tasks. Each vertebral disc from L5/S1 to T12/L1 in each scan is graded by an expert radiologist for the aforementioned gradings. In this setting, SCT takes as input volumes surrounding intervertebral discs (IVD) and outputs predictions for each grading. The study protocol for scans vary, though the vast majority of scans have at least a T1-weighted and T2-weighted sagittal scan. In this work we only consider sagittally sliced scans. However, there is no reason why SCT could not also operate on axial scans in conjunction. A more complete breakdown of the Genodisc dataset is given in [Jamaludin et al. 2017a]. We compare to SpineNet V1 [Jamaludin et al. 2017b] and SpineNet V2 [Windsor et al. 2022], existing models trained on the same dataset.

4.6 Experimental Results

This section describes experiments to evaluate the performance of SCT at the tasks of: (a) detecting spinal cancer; (b) grading common degenerative changes.

Preprocessing, Implementation & Training Details: For both tasks, vertebrae are detected and labelled in the scans using the automated method described in [Windsor et al. 2022]. This is then used to extract volumes surrounding VBs (for the cancer task) and IVDs (for the grading task). Volumes are resampled to $S \times 112 \times 112$ and $S \times 112 \times 224$ respectively, where S is the number of sagittal slices. Additional training details are given in the appendix.

Baselines: For each task, we compare SCT to baseline models operating on a single vertebra/IVD at a time. For the cancer task we compare to the visual encoder shown in Figure 4.1b operating alone on a single sequence (since this cannot be trivially extended to multiple sequences). For the radiological grading

task we compare with SpineNet V1 [Jamaludin et al. 2017b], SpineNet V2 [Windsor et al. 2022] and the visual encoder alone. We also train SCT on T1 and T2 scans independently.

Results: Table 4.2 shows the performance of all models at the cancer task. The SCT model performs well across all tasks. In particular, it performs much better at detecting subtle metastases (AUC: **0.80→0.931** for the expert labels). This effect can be seen clearly in Figure 4.3. This makes sense as patients will often present with multiple metastases and thus obvious metastases in one area of the spine will inform predictions on marginal cases elsewhere in the spine. We note slightly depreciated performance at the compression task. We believe this is due to overfitting as there are relatively few compression cases in the training set.

Table 4.3 shows the performance of all models at the radiological grading tasks, and compares to the state-of-the-art SpineNet model [Jamaludin et al. 2017a]. SCT outperforms SpineNet V1 & V2 on the same dataset for all tasks except central canal stenosis where the difference in performance is minimal . In total average performance increases from **85.9%→87.4**%. As expected, the multiple sequence model exceeds the performance of single sequence models in most tasks. The largest improvements can be seen in the endplate defect task (**82.9/87.8%** →**87.2/90.7%**), disc narrowing (**76.1%→77.4%**) and Pfirrmann grading (**71.0%** →**73.0%**). The Pfirrmann grading score of the T1-only model is far worse than the models with T2 sequences (64.5%). This matches expectations as one of the criteria of Pfirrmann grading is the intensity of the intervertebral discs in T2 sequences [Pfirrmann et al. 2001]. Overall, the improved performance is a clear indication of the benefit of considering context from multiple sequences and IVDs together.

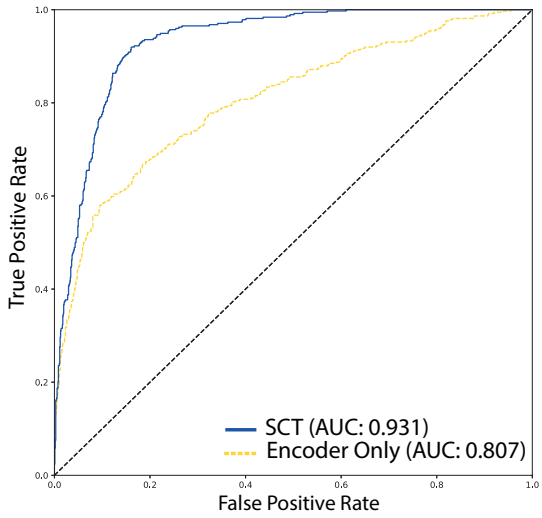


Figure 4.3: ROC curves for the metastases task for SCT and the encoder baseline. Curves for the other tasks are given in the appendix.

Table 4.2: AUC scores for the three tasks with various models. We compare to the report-extracted annotations and also expert annotations of each image.

Expert-Labelled Test Set Annotations

	ROC AUC		
	Mets	Frac.	Cmprs.
Baseline	0.80	0.975	**0.930**
SCT	**0.931**	**0.980**	0.868

Report-Extracted Test Set Annotations

	ROC AUC		
	Mets	Frac.	Cmprs.
Baseline	0.934	0.902	**0.955**
SCT	**0.944**	**0.901**	0.918

Model	# Classes:	Pfirrmann 5	Disc Narrowing 4	C.C.S. 2	Spondylolisthesis 2
SpineNet V1 (T2) [Jamaludin et al. 2017d]		71.0	76.1	**95.8**	95.4
SpineNet V2 (T2) [Windsor et al. 2022]		70.9	76.3	93.2†	95.0
Baseline: SCT Encoder (T2)		70.8	74.6	94.3	96.2
SCT (T1)		64.5	73.9	93.6	**96.6**
SCT (T2)		71.7	76.9	93.6	96.2
SCT (T1,T2)		**73.0**	**77.4**	94.9	95.5

Model	# Classes:	Endplate Defect Upper 2	Endplate Defect Lower 2	Marrow Change Upper 2	Marrow Change Lower 2	**Average**
SpineNet V1 (T2) [Jamaludin et al. 2017d]		82.9	87.8	89.2	88.4	85.8
SpineNet V2 (T2) [Windsor et al. 2022]		84.9	89.8	88.9	88.2	85.9
Baseline: SCT Encoder (T2)		87.2	88.2	90.0	89.2	86.3
SCT (T1)		84.7	87.0	88.4	88.7	84.7
SCT (T2)		85.3	89.3	**91.0**	**90.2**	86.8
SCT (T1,T2)		**87.2**	**90.7**	90.1	89.9	**87.4**

Table 4.3: Results of the IVD grading task. The balanced accuracy for each subtask is shown. CCS represents central canal stenosis. †For SpineNetV2, CCS is originally graded with 4 degrees of severity. We combine the severity classes 2-4 (mild, moderete & severe CCS) into a single class compared to class 1 (no CCS).

4.7 Conclusion

In this paper we present SpinalContextTransformer (SCT) for the analysis of multiple vertebrae and multiple MRI sequences. We demonstrate SCT being applied to a new task, detecting conditions related to spinal cancer including metastases, vertebral collapses and metastatic cord compression. We also show that the model improves on existing models for radiological grading of common spinal conditions and can be used flexibly with varying input imaging modalities. Finally, it is worth noting that the SCT architecture is applicable to the analysis of other modalities (e.g. CT, X-Ray), fields-of-view (e.g. coronal, axial) and to other repeated

anatomical sequential structures, such as teeth or ribs.

Acknowledgements and Ethics: Ethics for the spinal cancer dataset extraction are provided by OSCLMRIC (IRAS project ID: 207857). We are grateful to Dr. Sarim Ather, Dr. Jill Urban and Prof. Jeremy Fairbank for insightful conversations on the clinical aspect of this work as well as Prof. Ian McCall for annotating the data. Finally, we thank to our funders: Cancer Research UK via the EPSRC AIMS CDT and EPSRC Programme Grant Visual AI (EP/T025872/1).

4.8 Appendix

4.8.1 Pytorch-Style Pseudocode for SCT

```
class SpinalContextTransformer():
    def initialize(self, n_sequences, n_vertebrae):
        self.visual_encoder = VisualEncoderModel() # extracts features from each vertebra
        self.transformer    = TransfomerEncoderLayers(n_layers=2) # transforms features using context
        self.E              = 128 # embedding/token size for transformer
        # embedders convert one-hot encoded vert level/positions and type MR sequences
        # into something that can be summed with the features extracted from each vertebra
        self.positional_embedder = LinearLayer(in_dim=n_sequences, out_dim=self.E)
        self.sequence_embedder   = LinearLayer(in_dim=n_vertebrae, out_dim=self.E)

        self.sequence_attention = LinearLayer(in_dim=self.E, out_dim=self.E)
        # output classification layers
        self.met_classifier         = LinearLayer(in_dim=self.E, out_dim=1)
        self.fracture_classifier    = LinearLayer(in_dim=self.E, out_dim=1)
        self.compression_classifier = LinearLayer(in_dim=self.E, out_dim=1)

    def forward(self, vert_vols, one_hot_position_encoding, one_hot_sequence_encoding):
        B, N, C, S, H, W = vert_vols.shape
        B, N, C, n_sequences = one_hot_sequence_encoding.shape
        B, N, C, n_vertebrae = one_hot_position_encoding.shape

        # B=Batch size, N=Num Vertebrae, C=Num Channels/MR Sequences,
        # S=Number of Slices, H=Height, W=Width
        vert_vols = vert_vols.flatten(start_dim=0, end_dim=2)
        # [B,N,C,S,H,W] -> [B*N*C,S,H,W]
        vert_features = self.visual_encoder(vert_vols)
        # vert_features has dimension [B*N*C,E] where E is the embedding size (E=128 here)
        vert_features        = vert_features.reshape(B,N,C,S,self.E)
        positional_embedding = self.positional_embedder(one_hot_position_encoding)
        sequence_embedding   = self.sequence_embedder(one_hot_sequence_encoding)
        # add in positional and sequence embedding
        input_tokens  = vert_features + positional_embedding + sequence_embedding
        # used transformer to consider context between all vertebrae and sequences
        output_tokens = self.transformer(input_tokens)
        output_tokens = output_tokens.reshape(B,N,C,self.E)
        # now pool along MR sequence dimension (C)
        sequence_attention_scores = self.sequence_attention(output_tokens)
        # softmax across sequence dimension
        sequence_attention_weights = SoftMax(sequence_attention_scores,dim=2)
        pooled_output_tokens = (output_tokens*sequence_attention_weights).sum(dim=2)
        # pooled_output_tokens has dimension [B,N,E].
        # Now apply lineaer classifiers to get output of shape [B,N,1] for each task
```

```
        met_score      = self.met_classifier(pooled_output_tokens)
        fracture_score = self.fracture_classifier(pooled_output_tokens)
        mscc_score     = self.compression_classifier(pooled_output_tokens)
        return met_score, fracture_score, mscc_score
```

4.8.2 Additional Model Training Details

All models are trained until 10 successive validation epochs do not result in a decreased loss. An Adam optimizer is used with a learning rate of 10^{-4} and $\beta = (0.9, 0.999)$. A batch size of 20 is used for the lumbar-only degenerative changes task and 6 for the whole-spine cancer task (each sample consists of each VB in each MR sequence the scan). In total training for both tasks consumes around 40GB of GPU memory and takes approximately 6 hours using 2 Tesla P40s. The classification linear layers are then finetuned for each task for 10 epochs with the encoder and transformer weights frozen. Standard augmentations for each vertebra are used including rotation ($\pm 15°$), translation (± 32px), scaling ($\pm 10\%$) and intensity augmentation ($\pm 10\%$). For the multi-sequence models trained on T1 and T2 sequences, a single sequence or both sequences are dropped during training with probabilities of 0.4 and 0.1 respectively, leaving the model make predictions for a volume it has not seen based on context alone. 50% dropout is used for the transformer layers. Pytorch 1.10 was used to implement all models.

4.8.3 Slicewise Predictions

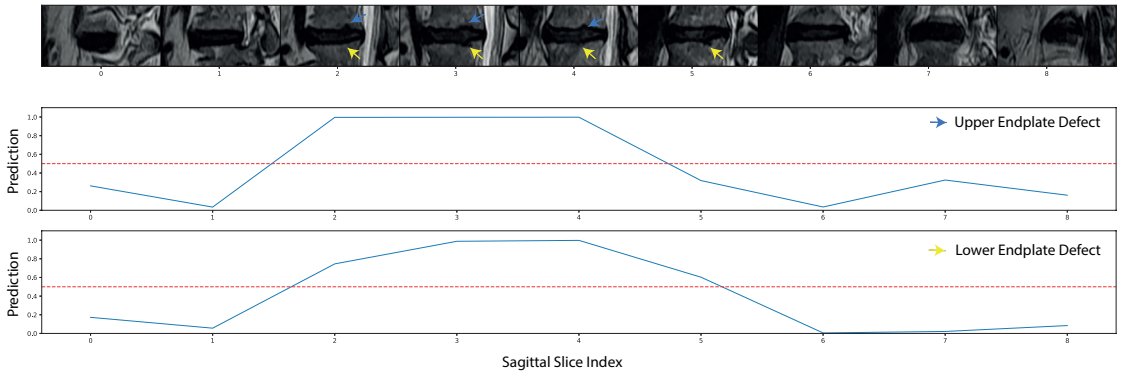


Figure 4.4: Slicewise outputs for SCT operating on an IVD (top row) with upper (blue arrows) and lower endplate defects (yellow arrows). The second and third rows show the predictions for upper and lower endplate defects respectively when each sagittal slice of the scan is fed into the model independently. SCT is highly accurate at detecting pathology even at the single slice level, detecting the upper endplate defects in slices 2,3 & 4 and the lower endplate is slices 2,3,4 & 5. This can considered a form of attribution. The dotted red line shows p(endplate defect) = 50% for both slicewise predictions.

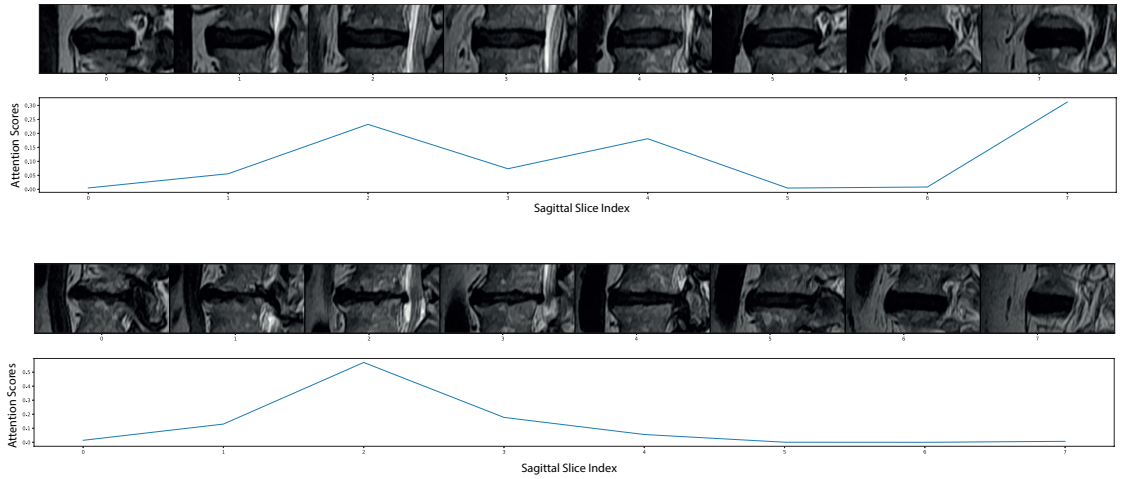


Figure 4.5: Slicewise attention values for a healthy (top) and an unhealthy (bottom) IVD. In the healthy IVD, attention is spread across all mid-sagital slices whereas in the unhealthy vertebra, the attention is concentrated on slices which clearly show pathology (endplate defects, narrowing and slight stenosis in slices 1,2 & 3).

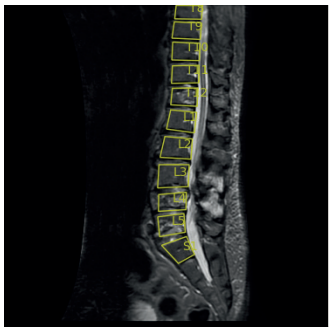
4.8.4 Example Annotation

Report:
Clinical Information: kidney cancer on follow up. CT scan shows stable disease. MRI spine to stage spine ? further progression

MRI Spine Whole : comparison is made with the MR scan dated [date redacted]
1. There has been a decrease in the size of the bone metastasis in L4. However, there has been further loss of vertebral body height at this level and there is protrusion of the posterior bony cortex into the spinal canal, with a slight increase in canal stenosis. There remains narrowing of the left exit foramen.
2. Abnormal bone marrow signal in L5 does not appear significantly changed.
3. The metastasis in T12 extending into the left posterior elements appears unchanged. There is no significant encroachment into the spinal canal at this level.
4. There is fatty replacement of marrow from T11-L5 in keeping with previous radiotherapy.
5. Right upper pole renal mass noted.
6. There is a large metastasis in the left side of the sacrum. This was incompletely imaged on the previous MR and cannot be compared.
7. No metastases are identified in the upper spine from C1 to T 11. No cord compression. There is a small metastasis in the clivus, unchanged.

Conclusion: Minor improvement in L4 metastasis but increase in spinal canal stenosis at this level due to a malignant compression fracture rather than tumour extension.

Scan Mid Sagittal Slices + Detected Vertebrae:



STIR Lumbar



STIR Cervical



T1 Lumbar

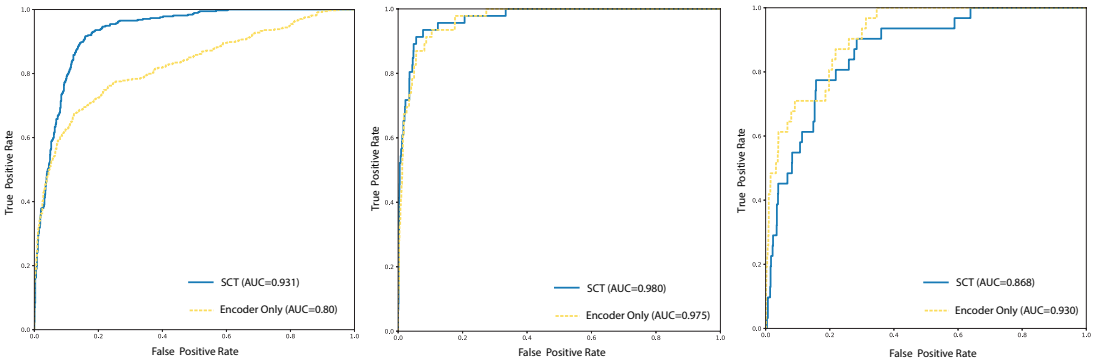
Extracted Annotations:

Condition	Widespread?	S1	L5	L4	L3	L2	L1	T12	T11	T10	T9
Metastasis?	✓			✓				✓			
Fracture?				✓							
Compression?				✓							

Condition	T8	T7	T6	T5	T4	T3	T2	T1	C7	C6	C5	C4	C3	C2
Metastasis?														
Fracture?														
Compression?														

Figure 4.6: An example of annotations extracted from a free-text report. There are metastases at T12 and L4, as well as cord compression and fractures ("malignant compression fracture") at L4. The report also suggests there maybe other metastases, e.g. in L5, however these are not explicitly stated to be present. For this reason the 'widespread' column is ticked for metastases. As a results all vertebrae other than T12 and L4 (i.e. levels where metastases are explicitly mentioned) are treated as 'unknown'.

4.8.5 ROC Curves for All Cancer Tasks



(a) Metastasis (b) Vertebral Fracture (c) Cord Compression

Figure 4.7: ROC Curves for the individual spine cancer tasks. These curves show the agreement of the model's predictions with the annotations produced by an expert spinal surgeon.

Part II

The Role of Self-Supervision In Radiological Imaging

Chapter 5

Self-Supervised Multi-Modal Alignment For Whole Body Medical Imaging

This chapter presents a paper published at MICCAI 2021, the 24$^{\text{th}}$ International Conference on Medical Image Computing and Computer Assisted Intervention (Strasbourg/Virtual).

Self-Supervised Multi-Modal Alignment For Whole Body Medical Imaging

Rhydian Windsor[1] Amir Jamaludin[1]

Timor Kadir[1,2] Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Enginnering Science, University of Oxford

[2]Plexalis

Abstract

This paper explores the use of self-supervised deep learning in medical imaging in cases where two scan modalities are available for the same subject. Specifically, we use a large publicly-available dataset of over 20,000 subjects from the UK Biobank with both whole body Dixon technique magnetic resonance (MR) scans and also dual-energy x-ray absorptiometry (DXA) scans. We make three contributions: (i) We introduce a multi-modal image-matching contrastive framework, that is able to learn to match different-modality scans of the same subject with high accuracy. (ii) Without any adaption, we show that the correspondences learnt during this contrastive training step can be used to perform automatic cross-modal scan registration in a completely unsupervised manner. (iii) Finally, we use these registrations to transfer segmentation maps from the DXA scans to the MR scans where they are used to train a network to segment anatomical regions without requiring ground-truth MR examples. To aid further research, our code will be made publicly available[§].

[§]`https://github.com/rwindsor1/biobank-self-supervised-alignment`

5.1 Introduction

A common difficulty in using deep learning for medical tasks is acquiring high-quality annotated datasets. There are several reasons for this: (1) using patient data requires ethical clearance, anonymisation and careful curation; (2) generating ground-truth labels may require expertise from clinicians whose time is limited and expensive; (3) clinical datasets are typically highly class-imbalanced with vastly more negative than positive examples. Thus acquiring sufficiently large datasets is often expensive, time-consuming, and frequently infeasible.

As such, there is great interest in developing machine learning methods to use medical data and annotations efficiently. Examples of successful previous approaches include aggressive data augmentation [Ronneberger et al. 2015] and generating synthetic images for training [Ghorbani et al. 2019]. Alternatively, one can use *self-supervised pre-training* to learn useful representations of data, reducing annotation requirements for downstream learning tasks. This method has already shown much success in other areas of machine learning such as natural image classification [T. Chen et al. 2020; Hénaff et al. 2020; He et al. 2020] and natural language processing [Mikolov et al. 2013; Devlin et al. 2019; Radford et al. 2019; Brown et al. 2020].

In this paper, we develop a self-supervised learning approach for cases where pairs of different-modality images corresponding to the same subject are available. We introduce a novel pre-training task, where a model must match together different-modality scans showing the same subject by comparing them in a joint, modality-invariant embedding space. If these modalities are substantially different in appearance, the network must learn semantic data representations to solve this problem.

In itself, this is an important task. Embeddings obtained from the trained networks allow us to check if two different scans show the same subject in large anonymised datasets (by verifying that their embeddings match). It also defines a notion of similarity between scans that has applications in population studies. However, the main reward of our method are the semantic *spatial* representations of the data learnt during training which can be leveraged for a range of downstream tasks. In this paper we demonstrate the embeddings can be used for unsupervised

rigid multi-modal scan registration, and cross-modal segmentation with opposite-modality annotations.

The layout of this paper is as follows: Section 5.2 describes the cross-modal matching task in detail, including the network architecture, loss function, and implementation details, as well as experimental results from a large, publically-available whole body scan dataset. Section 5.3 introduces algorithms using the embeddings learnt in Section 5.2 for fast unsupervised multi-modal scan registration which are shown to succeed in cases where conventional registration approaches fail. In Section 5.3.1, we then use these registrations to transfer segmentation maps between modalites, showing that by using the proposed cross-modal registration technique, anatomical annotations in DXAs can be used to train a segmentation network in MR scans.

5.1.1 Related Work

Self-supervised representation-learning is an incredibly active area of research at the moment. The current dominant praxis is to train models to perform challenging self-supervised learning tasks on a large dataset, and then fine-tune learnt representations for specific 'downstream' tasks using smaller, annotated datasets. Major successes have been reported in image classification [Asano et al. 2020; T. Chen et al. 2020; Hénaff et al. 2020; Grill et al. 2020; Caron et al. 2020], video understanding [Han et al. 2020; Qian et al. 2021] and NLP[Mikolov et al. 2013; Howard and Ruder 2018; Radford et al. 2019], with self-supervised approaches often matching or exceeding the performance of fully-supervised approaches.

Due to the existence of a few large, publically available datasets (such as [Johnson et al. 2019]), yet lack of large annotated datasets suitable for most medical tasks, self-supervised learning shows great promise in the medical domain. For example, previous work has shown it can be used to improve automated diagnosis of intervertebral disc degeneration [Jamaludin et al. 2017a] and common segmentation tasks [Taleb et al. 2020]. In [Taleb et al. 2021], it also is shown that using multiple MR sequences in self-supervised learning improves performance in brain tumour segmentation.

Data with multiple modalities is a natural candidate for self-supervised approaches,

as one can use information from one modality to predict information in the other. For example, previous work has shown self-supervised methods can benefit from fusion of the audio and visual streams available in natural video data [Alwassel et al. 2020; Arandjelović and Zisserman 2017; Arandjelović and Zisserman 2018; Owens and Efros 2018; Korbar et al. 2018]. In this paper we build on this multi-modal approach by extending it to explicit spatial registration across the modalities.

5.1.2 Dataset Information, Acquisition and Preparation

For the experiments in this paper we use data from the UK Biobank [Sudlow et al. 2015a], a large corpus of open-access medical data taken from over 500,000 volunteer participants. A wide variety of data is available, including data related to imaging, genetics and health-related outcomes. In this study we focus on two whole body imaging modalities collected by the Biobank: (1) 1.5T, 6-minute dual-echo Dixon protocol magnetic resonance (MR) scans showing the regions from approximately the neck to the knees of the participant with variation due to the subject's height and position in the scanner; (2) Dual energy x-ray absorptiometry (DXA) scans showing the entire body. In total, at the time of data collection, the Biobank consisted of 41,211 DXA scans and 44,830 MR scans from unique participants.

Our collected dataset consists of pairs of same-subject multi-sequence MR and DXA scans, examples of which can be seen in Figure 5.1. In total we find 20,360 such pairs. These are separated into training, validation and test sets with a 80/10/10% split (16,213, 2,027 and 2,028 scan pairs respectively). Scan pairs are constructed using (1) the fat-only and water-only sequences of the Dixon MR scans, and (2) the tissue and bone images from the DXA scans. For the purposes of this study, we synthesize 2D coronal images from the 3D MR scans by finding the mid-spinal coronal slice at each axial scan line by using the method described in [Windsor et al. 2020] to fit a curve to the spinal column. All scans are resampled to be isotropic and cropped to a consistent size for ease of batch processing (501×224 for MR scans and 800×300 for DXA scans). These dimensions maintain an equal pixel spacing of 2.2mm in both modalities. The scans are geometrically related in that the MRI field of view (FoV) is a cropped, translated and slightly

rotated transformation of the DXA scan's FoV. Both scans are acquired with the subject in a supine position, and there can be some arm and leg movements between the scans.

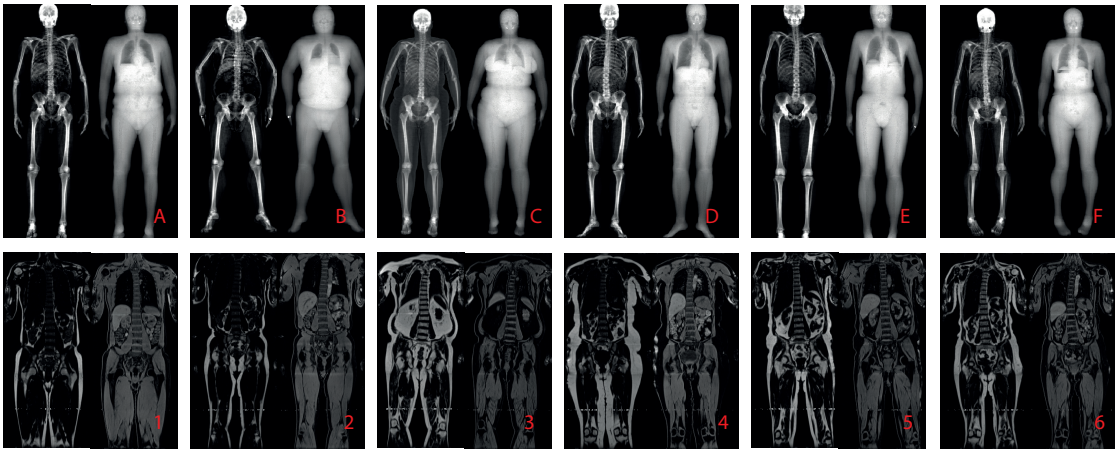


Figure 5.1: **Guess Who?** Example scan pairs from our dataset. The top row shows bone (left) and tissue (right) DXA scans from the dataset. The bottom row shows synthesized mid-coronal fat-only (left) and water-only (right) Dixon MR slices. In this paper, semantic spatial representations of the scans are learnt by matching corresponding DXA and MR scan pairs. Can you match these pairs?[1]

5.2 Matching Scans Across Modalities

This section describes the framework used to match same-subject scans across the DXA and MRI modalities. As shown in Figure 5.1, this is hard to perform manually with only a few scans. Differences in tissue types visible in DXA and MRI mean many salient points in one modality are not visible at all in the other. Furthermore, the corresponding scans are not aligned, with variation in subject position, pose and rotation.

To tackle this problem, we use the dual encoder framework shown in Figure 5.2, tasking it to determine the best alignment between the two scans such that similarity is higher for aligned same-subject scans than for aligned different-subject scans. Since both the DXA and MRI scans are coronal views and subject rotations relative to the scanner are very small, an approximate alignment requires determining a 2D translation between the scans. The similarity is then determined by a scalar product of the scans' spatial feature maps after alignment. In practice, this amounts to 2D convolution of the MRI's spatial feature map over the DXA's spatial feature map, and the maximum value of the resulting correlation map provides

a similarity score.

The network is trained end-to-end by Noise Contrastive Estimation [Gutmann and Hyvärinen 2012] over a batch of N randomly sampled matching pairs. If M_{ij} represents the similarity between the i^{th} DXA and j^{th} MRI, where $i = j$ is a matching pair and $i \neq j$ is non-matching, and τ is some temperature parameter, the total loss for the k-th matching pair, ℓ_k, is given by

$$\ell_k = -\left(\log \frac{\exp(M_{kk}/\tau)}{\sum_{j=1}^{N} \exp(M_{kj}/\tau)} + \log \frac{\exp(M_{kk}/\tau)}{\sum_{j=1}^{N} \exp(M_{jk}/\tau)}\right) \tag{5.1}$$

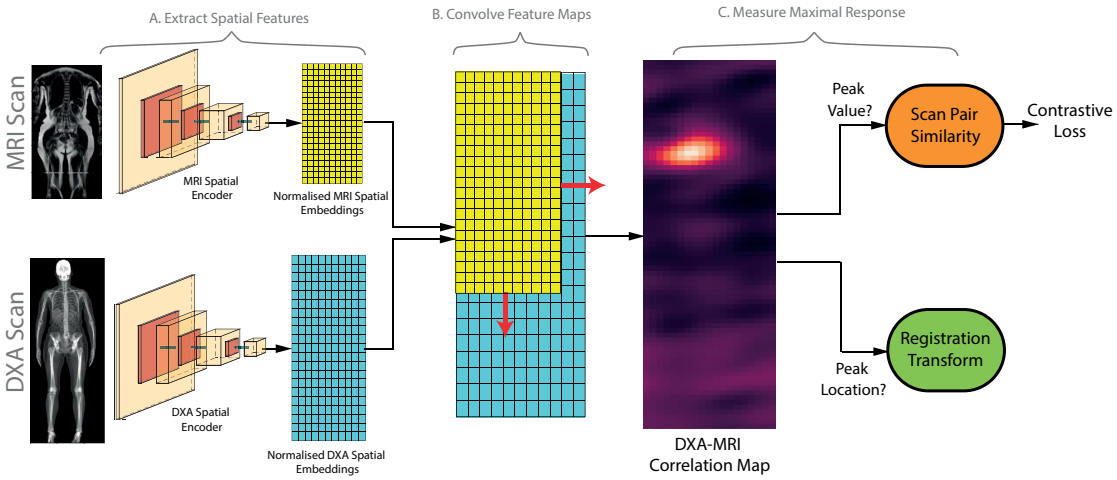


Figure 5.2: The dual encoding configuration used for contrastive training. Two CNNs ingest scans of the respective modalities, outputting coarse spatial feature maps (A). The feature maps of each DXA-MRI pair are normalised and correlated to find the best registration (B). Using this registration, the maximum correlation is recorded as the similarity between the two scans (C). The architecture used for both spatial encoders is shown in the appendix.

5.2.1 Experiments

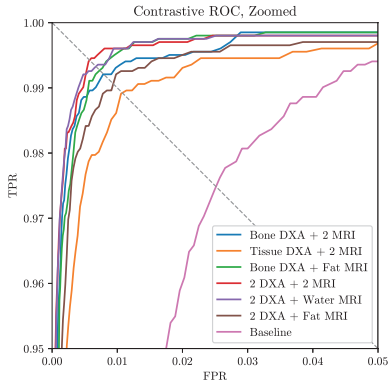
This section evaluates the performance of the proposed configuration on the cross-modal scan-matching task. To determine the relative importance of each MRI sequence and each DXA type, we train networks varying input channels to each modality's encoder (see Figure 5.3 for the configurations used). To demonstrate the value of comparing spatial feature maps of scans instead of a single global embedding vector, we compare to a baseline network that simply pools the spatial

[1] $A \to 5$, $B \to 3$, $C \to 4$, $D \to 2$, $E \to 1$, $F \to 6$

feature maps into a scan-level descriptor, and is trained by the same contrastive method. Details of this baseline are given in the appendix.

Implementation. Networks are trained with a batch size of 10 using an Adam optimizer with a learning rate of 10^{-5} and $\beta = (0.9, 0.999)$. A cross-entropy temperature of $T = 0.01$ is used (a study of the effect of varying this is given in the appendix). Spatial embeddings are 128-dimensional. Training augmentation randomly translates the both scans by ± 5 pixels in both axis and alters brightness and contrast by $\pm 20\%$. Each model takes 3 days to train on a 24GB NVIDIA Tesla P40 GPU. Networks are implemented in PyTorch v.1.7.0.

Evaluation measures. We evaluate the quality of the learnt embeddings on the test set by assessing the ability of the system to: (1) retrieve the matching opposite modality scan for a given query scan based on similarity; (2) verify if a given scan pair is matching or not. In the latter case, positive matches to a scan are defined as those with similarities above a threshold, ϕ, and negative matches have similarity $\leq \phi$. Considering all possible DXA-MRI scan pairs (matching & non-matching), we can then generate an ROC curve by varying ϕ from -1 to 1. For the retrieval task, we report top-1 and top-10 recall based on similarity across all test set subjects, and the mean rank of matching pairs. For the verification task, we report the ROC area-under-curve (AUC) and the equal error rate (EER) (i.e. when $TPR = FPR$).



(a) ROC Curve

Input		Verification		Retrieval		
DXA	MRI	AUC	EER (%)	% Recall @1	% Recall @10	Mean Rank
Baseline		0.9952	2.57	26.3	78.7	9.246
B	F	0.9992	0.77	89.4	99.4	2.106
B	F,W	**0.9993**	0.84	87.7	99.4	2.079
T	F,W	0.9986	1.14	83.1	98.4	3.013
B,T	F	0.9989	0.98	85.8	98.7	2.569
B,T	W	**0.9993**	0.70	90.1	99.4	**1.920**
B,T	F,W	0.9992	**0.60**	**90.7**	**99.5**	2.526

(b) Retrieval and Verification Performance

Figure 5.3: Verification and retrieval performance on the 2,028 scan test dataset with varying inputs of bone DXA (B), tissue DXA (T), fat-only MR (F), and water-only MR (W). (c) shows an ROC curve for the verification task. Table (b) reports performance statistics for the models, including equal error rate (EER), area under curve (AUC), recall at ranks 1 & 10 and the mean rank of matches.

Results. Figure 5.3 shows the ROC curve and performance measures for varying

input channels. All configurations vastly exceed the baseline's performance, indicating the benefit of spatial scan embeddings as opposed to scan-level descriptor vectors. The full model achieves a top-1 recall of over 90% from 2028 test cases. The tissue DXA-only model performs worst of all configurations suggesting bone DXAs are much more informative here. Extended results and recall at K curves are given in the appendix.

Discussion. The strong performance of the proposed method on the retrieval task by matching spatial (as opposed to global) features is significant; it suggests the encoders learn useful semantic information about specific regions of both scans. This has several possible applications. For example, one could select a query ROI in a scan, perhaps containing unusual pathology, calculate its spatial embeddings and find similar examples across a large dataset (see [Simonyan et al. 2011] for a more detailed discussion of this application). More conventionally, the learnt features could be also used for network initialization in downstream tasks on other smaller datasets of the same modality, potentially increasing performance and data efficiency. As a demonstration of the usefulness of the learnt features, the next section of this paper explores using them to register scans in a completely unsupervised manner.

5.3 Unsupervised Registration Of Multi-Modal Scans

A major advantage of this contrastive training method is that dense correspondences between multi-modal scans are learnt in a completely self-supervised manner. This is non-trivial; different tissues are shown in each modality, making intensity-based approaches for same- or similar-modality registration [Viola and Wells 1995; Lowe 2004] ineffective. Here we explore this idea further, developing three methods for estimating rigid registrations between the modalities. Each method is assessed by measuring L2-distance error when transforming anatomical keypoints from the MRIs to the DXA scans. For each proposed registration method the aim is to estimate the three transformation parameters; a 2D translation and a rotation.

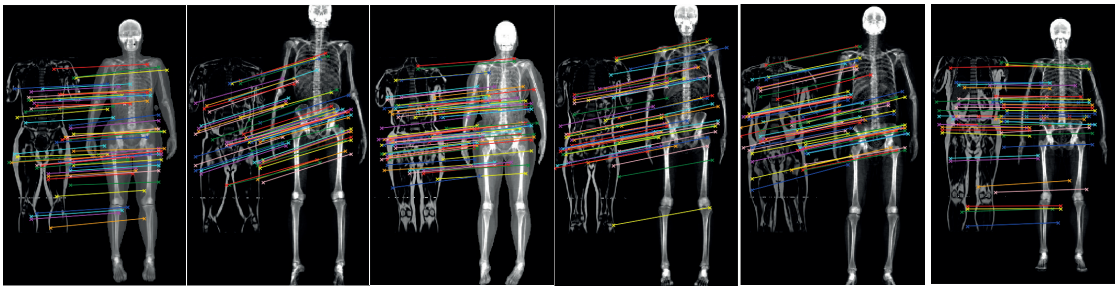


Figure 5.4: Salient point correspondences between scan pairs found by Lowe's ratio test & RANSAC. The fat-only channel of MRI source image is shown on the left, with the target DXA bone scan shown on the right

1. Dense Correspondences: During training, the contrastive framework attempts to align dense spatial feature maps before comparing them. We can use this to determine the registration translation by convolving the feature maps together and measuring the point of maximum response as the displacement between the images (as in Figure 5.2, stages A, B). The rotation between the scans is found by rotating the MRI scan across a small range of angles, convolving the feature maps, and recording the angle which induces the greatest alignment.

2. Salient Point Matching: The dense correspondence method is slow, especially on a CPU, as it requires multiple convolution operations with large kernels. To speed up registration we need use only a few salient points between the feature maps. These can be found by matching pairs of points based on correlations and then employing Lowe's second nearest neighbour ratio test [Lowe 1999] to remove ambiguous correspondences, followed by RANSAC estimation of the transformation. Details of this algorithm are given in the appendix. Example correspondences found by this method are shown in Figure 5.4.

3. Refinement Regressor: The previous approaches generate robust approximate registrations between the two images but are limited by the resolution of the feature maps they compare (8×downsample of a 2.2mm pixel spacing original image). To rectify this issue we use a small regression network to refine predictions by taking the almost-aligned feature maps predicted by the aforementioned methods and then outputting a small refinement transformation. Training data for this task can be generated 'for free' by taking aligned scan pairs from the salient point matching method and slightly misaligning them with a randomly sampled rotation and translation and then training a network to regress this random transformation. The regression network is trained on 50 aligned pairs predicted by the

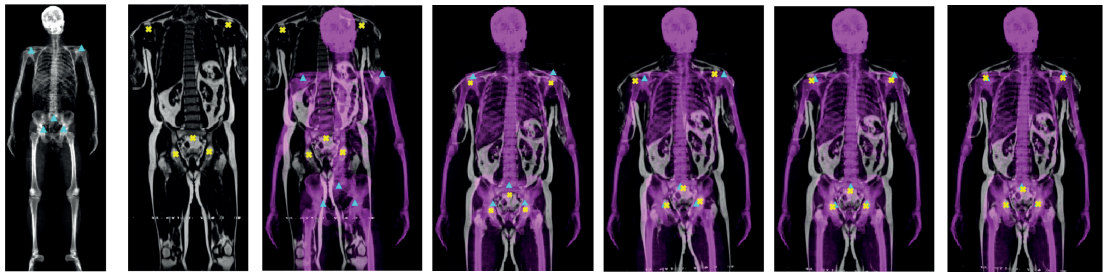


Figure 5.5: Example results from each registration method. (a) & (b) show keypoints for the MRI and the DXA. The MRI & keypoints are registered to the DXA by: (c) no transform; (d) mutual information maximisation; (e) Dense correspondences as in pre-training; (f) Salient point matching via Lowe's ratio test; (g) Applying the refinement regressor to (f).

previous salient point matching method and manually checked for accuracy. These are not expected to be perfect registrations, but should be close enough to allow the refinement network to learn how to correct erroneously aligned scans to at least an acceptable registration. For each pair, several copies are generated with slight randomly sampled translations and rotations at the original pixel resolution. For each transformed pair, the DXA and MRI spatial feature maps are then concatenated together, and used as input to a small CNN followed by a fully-connected network that estimates the three parameters of the transformation for each pair.

Experiments. To measure the quality of these registrations, 5 keypoints were marked in 100 test scan pairs: the femoral head in both legs (hip joints), humerus head in both arms (shoulder joints) and the S1 vertebra (base of the spine). MRI keypoints are annotated in 3D and then projected into 2D. These keypoints provide the ground truth for assessing the predicted transformations. Example annotations are shown in Figure 5.5. We then measure the mean L2-localisation error when transferring the keypoints between modalities using rigid body transforms predicted by the proposed methods. We compare to baselines of (i) no transformation (i.e. the identity); and (ii) rigid mutual information maximisation[2]. To measure annotation consistency and error induced by change in subject pose, we also report the error of the 'best-possible' rigid transformation keypoints - that which minimises the mean L2 transfer error.

Results. Table 5.1 shows the localisation error achieved by each method. All

[2]Using MATLAB's `imregister` with `MattesMutualInformation`[Mattes et al. 2001] as an objective.

methods yield accurate registrations between the images. The best method is found to be salient point matching followed by the refinement network, which is also shown to be fast on both a GPU and CPU. We attempted to calculate SIFT and MIND features in both modalities and match them as proposed in [Toews et al. 2013] and [Heinrich et al. 2012] repectively however these approaches did not work in this case (see appendix).

Discussion. In this setting, our methods were found to outperform other approaches for multi-modal registration (mutual information, MIND and SIFT). We believe the reason for this is that DXA scans show mostly bony structures, whereas most visual content in MRI is due to soft tissues which can't be differentiated by DXA. As such, most pixels have no obvious intensity relation between scans. However, accurate registration between the scans is important as it allows collation of spatial information from both modalities. This can be exploited in at least two ways: (i) for joint features; registration allows shallow fusion of spatial features from both modalities. This could be useful in, for example, body composition analysis, conventionally done by DXA but which may benefit from the superior soft tissue contrast of MRI [Borga 2018]. (ii) for cross-modal supervision; registration allows prediction of dense labels from one modality which can then be used as a training target for the other. For example one could diagnose osteoporosis/ fracture risk at a given vertebral level from MR using labels extracted from DXA by conventional methods.

Method	Keypoint Transfer Error (cm)					Time(s)	
	HJ	S1	SJ	Median(all)	Mean(all)	GPU	CPU
No Trans.	22.1±5.0	21.7±5.3	22.3±5.0	21.9	22.01±5.1	0	0
Mut. Inf.	2.23±1.3	2.67±1.4	2.75±2.2	2.21	2.52±1.7	-	1.0
Dense Corr.	1.48±0.8	1.52±0.8	2.05±1.2	1.52	1.72±1.0	1.5	5.7
Sal. Pt. Mt.	1.34±0.9	1.37±1.0	2.04±1.4	1.46	1.63±1.3	0.4	1.1
Regressor	1.24±0.8	1.30±0.9	1.44±0.9	1.12	1.32±0.9	0.9	1.5
Best Poss.	0.84±0.4	0.84±0.5	0.87±0.4	0.84	0.87±0.4	-	-

Table 5.1: Keypoint transfer error for the proposed methods. We report the mean and median error for all keypoints combined and for the hip joints (HJ), shoulder joints (SJ) and S1 individually. Runtime on a GPU & CPU is also shown.

5.3.1 Cross-Modal Annotation Transfer

A benefit of the demonstrated cross-modal registrations is that they allow the transfer of segmentations between significantly different modalities, meaning seg-

mentation networks can be trained in both modalities from a single annotation set. This is useful in cases when a tissue is clearly visible in one modality but not the other. For example, here the pelvis is clearly visible in the DXA scan but not in the MRI slice. As an example of using cross-modal annotation transfer, the spine, pelvis and pelvic cavity are segmentated in DXA scans using the method from [Jamaludin et al. 2018]. These segmentations are then transferred to the MRI scans by the refinement network from section 5.3. where they act as pixel-wise annotations to train a 2D U-Net [Ronneberger et al. 2015] segmentation network. Compared to manual segmentation of the spine performed in 50 3D MR scans and projected into 2D, this network achieves good performance, with a mean Dice score of 0.927 showing the quality of the transferred annotations. Examples are shown in the appendix.

5.4 Conclusion

This paper explores a new self-supervised task of matching different-modality, same-subject whole-body scans. Our method to achieves this by jointly aligning and comparing scan spatial embeddings via noise contrastive estimation. On a test dataset of 2028 scan pairs our method is shown to perform exceptionally well with over 90% top-1 recall. We then show the learnt spatial embeddings can be used for unsupervised multi-modal registration in cases where standard approaches fail. These registrations can then be used to perform cross-modal annotation transfer, using DXA segmentations to train a MRI-specific model to segment anatomical structures. Future work will explore using the learnt spatial embeddings for other downstream tasks and extend this method to 3D scans.

Acknowledgements. Rhydian Windsor is supported by Cancer Research UK as part of the EPSRC CDT in Autonomous Intelligent Machines and Systems (EP/L015897/1). We are also grateful for support from a Royal Society Research Professorship and EPSRC Programme Grant Visual AI (EP/T028572/1).

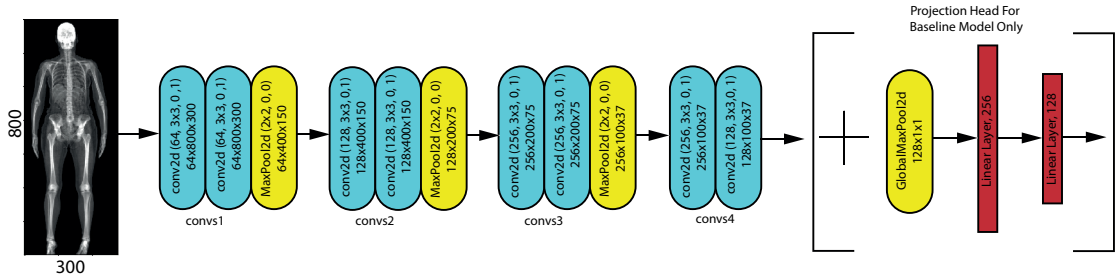


Figure 5.6: The simple spatial encoder used in the contrastive framework. Both the MR and DXA spatial encoders use this architecture. For the baseline model, the output spatial features are max pooled and the projection head shown on the right is appended. Each convolutional & linear layer except the final one uses ReLU activations, followed by BatchNorm.

5.5 Appendix

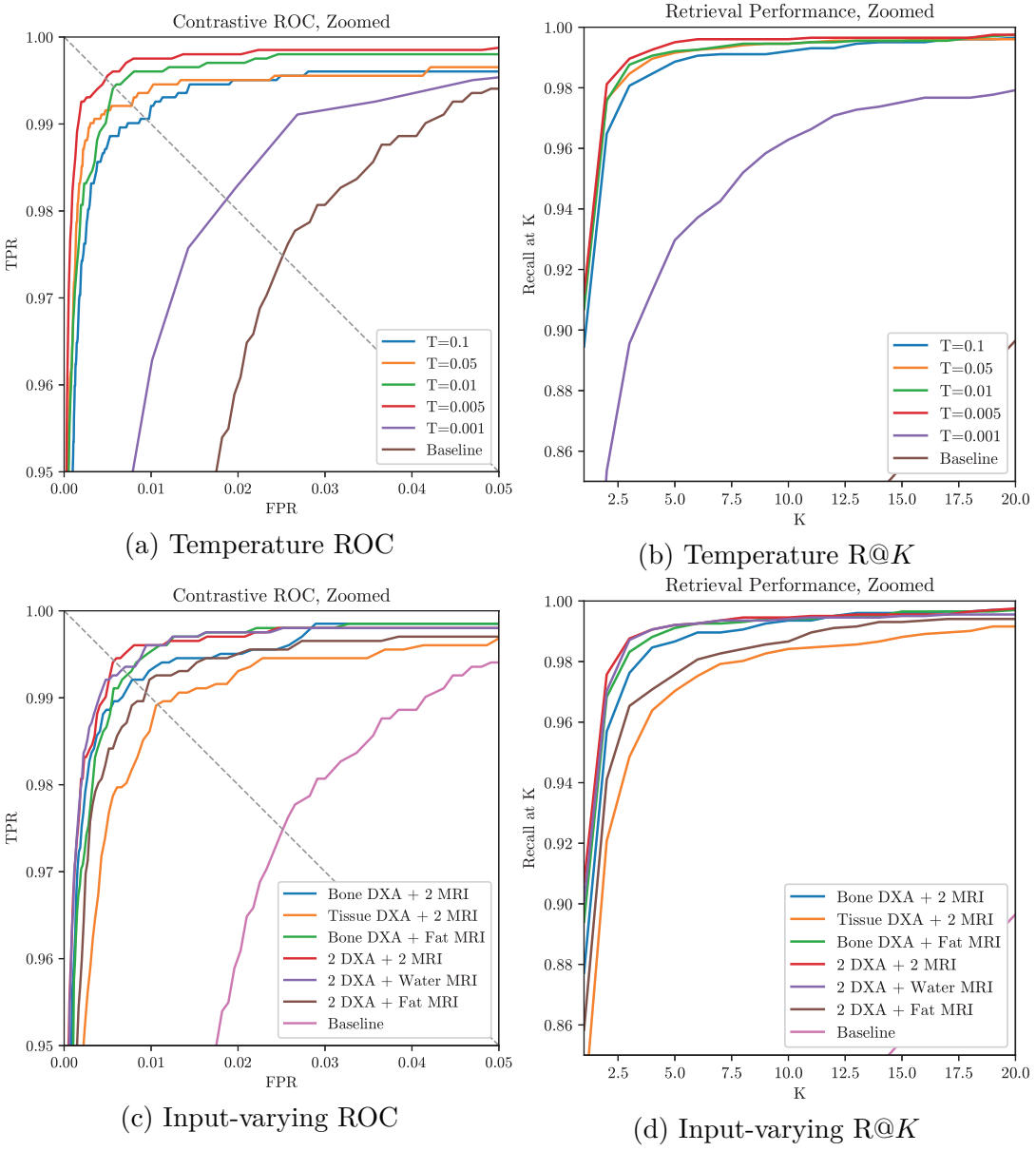
5.5.1 Scan Matching - Architecture, Extended Tables & Figures

Input Scans	% Recall Top-1	Top-5	Top-10	AUC	Mean Rank	Equal Error Rate	TPR@ FPR=1%
Bone DXA + Fat MRI	89.41	99.11	99.41	0.9992	2.106	0.0077	0.9955
Bone DXA + 2 MRI	87.72	98.66	99.36	**0.9993**	2.079	0.0084	0.9935
Tissue DXA + 2 MRI	83.12	97.03	98.42	0.9986	3.013	0.0114	0.9891
2 DXA + Fat MRI	85.84	97.57	98.66	0.9989	2.569	0.0098	0.9925
2 DXA + Water MRI	90.05	**99.21**	99.41	0.9993	**1.920**	0.0070	**0.9960**
2 DXA + 2 MRI	**90.69**	**99.21**	**99.46**	0.9992	2.526	**0.0060**	**0.9960**

Table 5.2: An extended table of performance metrics for varying scan-input in contrastive training including true positive rate at a false positive rate of 1% (TPR@FPR=1%) and top-5 recall.

Temperature, τ	% Recall Top-1	Top-5	Top-10	AUC	Mean Rank	Equal Error Rate	TPR@ FPR=1%
τ=0.1	89.46	98.86	99.41	0.9991	2.148	0.0095	0.9921
τ=0.05	91.14	99.16	99.45	0.9991	**2.140**	0.0080	0.9946
τ=0.01	90.07	99.21	99.45	0.9992	2.527	0.0060	0.9960
τ=0.005	**91.18**	**99.50**	**99.60**	**0.9994**	2.214	**0.0049**	**0.9975**
τ=0.001	74.41	92.97	96.28	0.9979	3.534	0.0197	**0.9629**

Table 5.3: Scan-matching performance metrics for configurations with varying softmax temperature, τ.



(a) Temperature ROC

(b) Temperature R@K

(c) Input-varying ROC

(d) Input-varying R@K

Figure 5.7: ROC and Recall at K (R@K) curves for varying scan-input and temperature parameter τ.

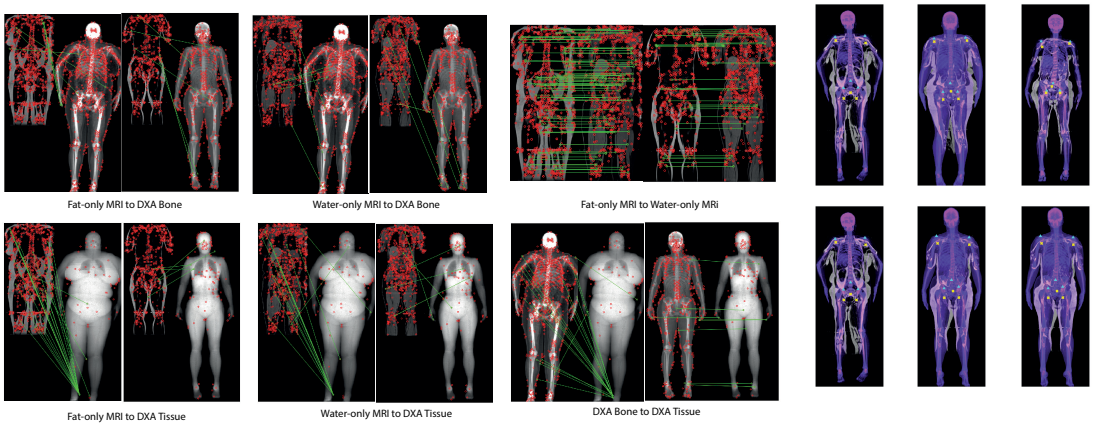
5.5.2 Unsupervised Registration

Lowe's Nearest Neighbours Ratio Test

1. For pixel in source feature map s_1, find the top-two most correlating pixels in the target feature map, t_1 and t_2 respectively.
2. If $\tau \cdot sim(s_1, t_1) < sim(s_1, t_2)$ save the pair (s_1, t_1), where τ is some threshold between 0 and 1 and sim is the cosine similarity.
3. Repeat this for each pixel in the source feature map to obtain a set of candidate matches between the feature maps.

4. Apply RANSAC to remove spurious correlations from these candidates

5. Use LMEDS to get the best rigid transform between remaining inlying points.



(a) SIFT Correspondences

(b) Non-rigid MIND Registration

Figure 5.8: **Attempted registration using SIFT & MIND features for the varying modalities**. a) SIFT features (shown by red circles) were calculated in both the original image and a negative version. They are then matched across modalities by brute-force matching and RANSAC is applied to find the best affine transform between the images. The in-lying matches are shown in green. This approach only succeeds finding correspondances between the already aligned MR sequences and to, some extent, the DXA images. b) Results from Gauss-Newton optimised non-rigid MIND registration results as implemented at `https://github.com/cmirfin/BBR`.

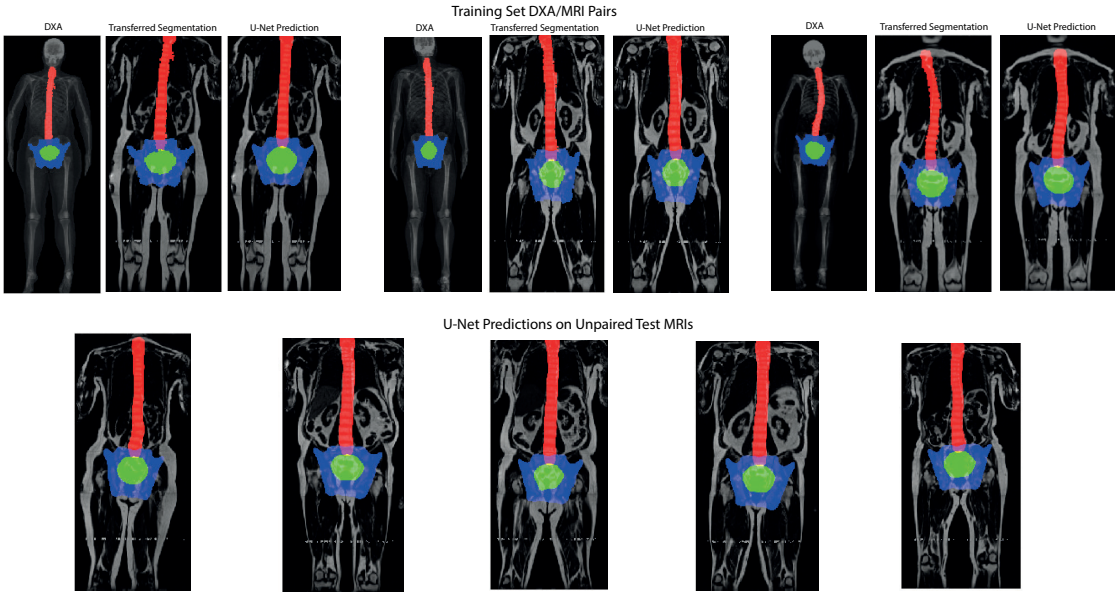


Figure 5.9: **Predicted segmentations of the spine, pelvis and pelvic cavity in MR scans by a U-Net trained with DXA annotations**. Structures are segmented in DXA scans and transferred to the corresponding MR scan by the refinement registration method. A model is trained on the transferred segmentations which can then be applied to unpaired MR scans.

Chapter 6

3D Spinal Column Segmentation with Single Plane 2D-Projected Annotations

This chapter presents a paper published at the Medical Imaging Meets Neurips Workshop at Neurips 2021 (Virtual).

3D Spinal Column Segmentation with Single Plane 2D-Projected Annotations

Rhydian Windsor[1] Amir Jamaludin[1]

Timor Kadir[1,2] Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Enginnering Science, University of Oxford

[2]Plexalis

Abstract

In this work, we explore methods of performing 3D segmentation using single-plane 2D projections of target segmentation maps for supervision. Specifically, we attempt to segment the spine in 3D using annotations derived from registered 2D coronal-plane scans of the same patient. We propose several methods to perform this task by exploiting prior knowledge of the spine's 3D shape and appearance. These methods are tested empirically using DXA-MRI scan pairs from the UK Biobank. The best-performing segmentation model achieves good agreement with manual annotations of the spine, with a 3D Dice score of 0.642. Performing this segmentation allows estimation of the 3D spinal curve, information which has previously been shown to improve monitoring of spinal scoliosis.

6.1 Introduction

This work investigates segmenting 3D structures using 2D projections of target segmentations in a single plane of view. This challenge is interesting for several reasons. Firstly, it significantly reduces annotation burden as annotating in 2D is much faster and more precise than in 3D. This concept has already been explored in training segmentation models based on supervision for only some slices of a 3D

image [Çiçek et al. 2016]. The methods presented here can be considered as an extreme version of this, where only a single slice is annotated.

Our method also allows information readily available from 2D planar scans such as X-rays or DXA scans to train models which operate on 3D modalities such as MRI or CT. Here, we draw on a problem from this setting. We aim to use existing methods which segment the spine in 2D coronal DXA scans to train a model to segment the spine in 3D in MRI scans. Determining the 3D curve of the spine is clinically important. For example, scoliosis is a 3D problem which frequently requires surgical intervention. However, assessment is typically done using 2D planar scans such as X-rays or DXA scans [Jamaludin et al. 2020a]. Such scans only give information about the 2D-projected shape of the spine, and thus could lead to erroneous clinical decision making. Existing work has shown that, by knowing the spine's 3D shape, we can improve predictions on the deformity development and surgery outcome [Pasha and Flynn 2018].

3D segmentation from 2D projections of targets is a difficult task, especially when all target projections are in a single plane of view. The main challenge is to discern which voxels are part of the object we are intending to segment and which belong to other structures 'in line' with the target object when the segmentation map is projected into 2D. Since all projections are in the same plane, the same structures will always be marked as positives in the target projection. Hence a key question is how to constrain the predictions in the projection axis so that only correct voxels are marked as foreground.

6.2 3D Segmentation with 2D Annotations

The general setup for the segmentation experiments in this work is shown in Figure 6.1. With a Dixon MR scan as input, a 3D UNet predicts a volumetric segmentation of the spine. This 3D segmentation map is then maximally projected into the coronal plane where it is compared to a target 2D-coronal projection via a standard 2D segmentation loss, such as the Dice loss.

The challenge here is clear; for each row of voxels in the coronal/frontal axis the model must solve a multiple instance learning problem, determining which

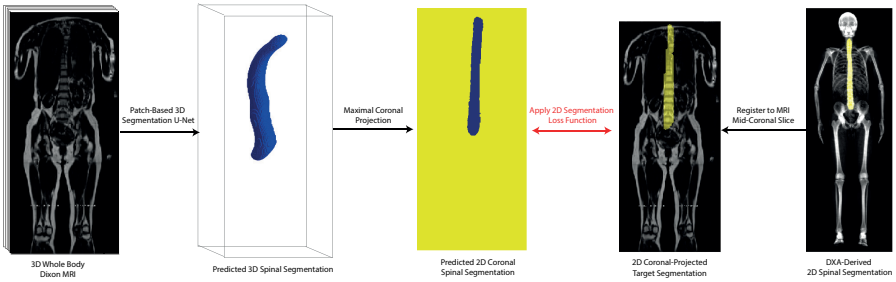


Figure 6.1: The framework used to train 3D spinal column segmentation networks in this paper.

voxel(s) correspond to the vertebral bodies and which do not. This challenge is compounded by the fact that voxels belonging to structures directly behind or in front of the scan will be in every positively-labelled bag (a row of voxels in the coronal axis). The following section suggests methods to solve this problem by using prior knowledge of the spine's shape and structure.

Coronal Sparsity: A simple idea is to use a loss function which encourages predicted segmentation maps to be sparse in the coronal axis. This can be done by adding an auxiliary term to the 2D segmentation loss function, $\mathcal{L}_{seg}$, which minimises the L1-norm of each row of coronal voxels in which the 2D target segmentation tells us the spine is present;

$$\mathcal{L}(\hat{\mathbf{Y}}, Y) = \mathcal{L}_{seg.}(\hat{Y}, Y) + \frac{\beta}{\sum_{i,j} \mathbb{1}_{ij}} \sum_{i,j} \mathbb{1}_{ij} ||\hat{\mathbf{y}}_{ij}||_1 \tag{6.1}$$

where $\hat{\mathbf{Y}}$ is the predicted 3D spinal segmentation, $\hat{Y}$ and Y are the coronally-projected prediction and target segmentation respectively and $\hat{\mathbf{y}}_{ij}$ is the row of voxels in the coronal axis at location (i, j) in the coronal plane. $\mathbb{1}_{ij}$ is an indicator function which is 1 if the projected target is positive in the coronal voxel row at (i, j) or 0 otherwise. β is a parameter to balance the magnitude of this term with $\mathcal{L}_{seg.}$.

Width-to-depth ratio consistency: A stronger assumption is to note that the sagittal width of vertebral bodies is approximately the same as their coronal depth (the average width-to-depth ratio is 1.38 in modern 46 year-old adults [Korpinen et al. 2020]). Therefore, we can introduce an additional term, $\mathcal{L}_{shape}$, to the loss function encouraging this to be true for predicted segmentations. In this work,

we use the form $\mathcal{L}_{shape} = \gamma \sum_i |w_i - \hat{d}_i|$, where i indexes the axial slice/height, w_i is the maximum sagittal width of the target segmentation in this slice and $\hat{d}_i$ is the predicted coronal depth. $\hat{d}_i$ is calculated by maximally projecting the predicted 3D segmentation map into the sagittal plane and summing the voxel-wise segmentation probabilities along coronal axis. γ balances $\mathcal{L}_{shape}$ with $\mathcal{L}_{seg.}$.

Symmetric Filters: Disregarding the pedicle protrusions, vertebral bodies appear as approximate elliptic cyclinders in 3D MR scans. As such they have approximate rotational symmetry of order 4 about the axial axis (i.e. rotating a vertebral body by multiples of 90° about this axis does not change its appearance from any perspective). They are also roughly symmetric to flips in the sagittal and coronal axes. Therefore, a network with filters which have this symmetry and can accurately determine the sagittal extent of a vertebral body can also likely accurately determine its coronal extent. Symmetrical filters could be enforced by using convolutional kernels that display these symmetries. However, in this work we encourage this symmetry by augmentation, randomly flipping the input volume in the sagittal and/or coronal axis and also rotating about the axial axis by random multiples of 90°before feeding it to the network. The output is then flipped/rotated back to the input's original orientation before applying the loss function.

6.3 Experiments

Data: Our dataset consists of whole body images of two modalities from the UK Biobank [Sudlow et al. 2015b]. These are: a) 3D Dixon technique MRI scans (we use the water-suppressed sequence in this work); b) 2D coronal DXA bone scans. At the time of data access, the UK Biobank consisted of 20,268 subjects with both a DXA and MRI scan. This is split into a training, validation and testing sets of size 16213, 2027 and 2028 scan pairs respectively. From the test and validation sets, 50 MRIs are selected at random and all vertebral bodies are annotated by a non-expert. For each DXA in the training dataset the spine is segmented using the automated method outlined in [Jamaludin et al. 2019] and registered to the mid-coronal slice of the MRI using the Lowes' ratio test-derived method from [Windsor

et al. 2021].

Training: For all segmentation models we use a 3D U-Net [Çiçek et al. 2016] architecture with Instance Norm, kernels of size 3×3×3, and dilation of size 1×1×1 in each downsampling block. The architecture operates on patches of 135×156×224 with an isotropic resolution of 2.23mm^3 per voxel. Four models are trained: a) a baseline method which simply applies the 2D Dice loss to the predicted 3D segmentation projected into the coronal plane; b) a method that uses this same loss function with the auxiliary coronal sparsity constraint described in Section 6.2 ($\beta = 10^{-3}$); c) the baseline model with an additional shape constraint term in the loss function ($\gamma = 10^{-2}$); d) the baseline model using the flipping and rotation augmentations described in the final paragraph of Section 6.2. Each model is trained using an Adam optimiser with learning rate 10^{-3}, $\beta = (0.9, 0.999)$, and a batch size of 6 for 48 hours. Models are validated every 100 iterations.

Evaluation: After training, the models are evaluated on the 50 withheld manual 3D spine annotations from the test dataset. We report the 3D Dice coefficient as well as the 2D Dice coefficient when the predicted and target 3D segmentation maps are maximally projected into the coronal and sagittal planes. These results are given in Table 6.1. Additionally, Figure 6.2 shows qualitative results of each method on a sample from the test dataset.

Discussion: As can be seen from Table 6.1, the best method suggested here is the symmetric filters method. Enforcing sparsity and consistency in the predicted segmentation's width-to-depth ratio improved 3D Dice results slightly over the ground truth method however resulted in segmentations of mostly the back of the vertebrae (the pedicles) rather than the vertebral bodies themselves. Future work will focus on using segmentations derived from methods such as these to estimate the 3D curve of the spine in MRI which has a range of clinical benefits as evidenced in [Pasha and Flynn 2018]. Another possible area of research is using the 3D curve found by this method to predict parameters such as torsion on the spine from 2D coronal planar scans such as DXA scans.

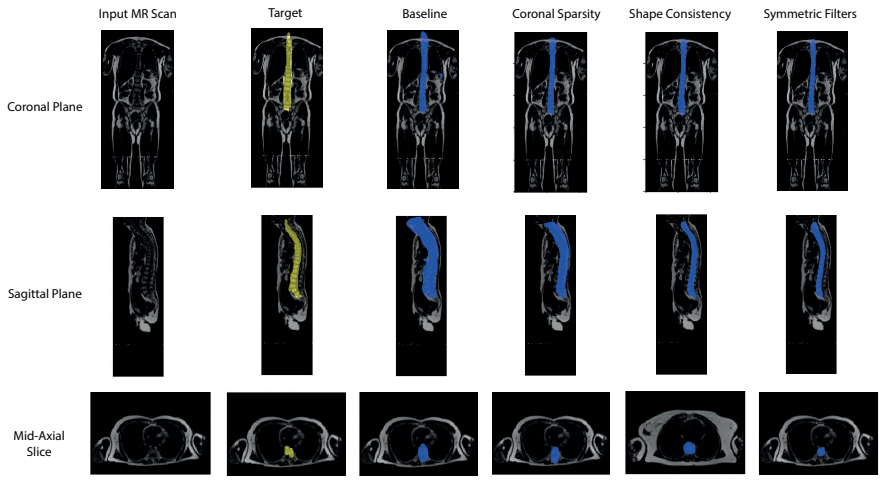


Figure 6.2: Example 2D projected segmentations by each suggested method compared with the target 3D segmentation.

Method	3D Dice	2D Projected Dice	
		Coronal	Sagittal
Baseline	0.356	0.826	0.482
Coronal Sparsity	0.427	0.828	0.585
Shape Consistency	0.452	0.829	0.594
Symmetric Filters	0.642	0.827	0.779

Table 6.1: Segmentation results for the proposed methods. We report the 3D Dice coefficient as well as the 2D Dice score when segmentations are maximally-projected coronally and sagitally respectively.

6.4 Broader Impact Statement

This work and [Jamaludin et al. 2019] suggest that spinal curvature can be measured using both DXA and MRI scans instead of traditional planar X-rays. This has potential benefits to patients. DXA scans are lower energy that traditional X-rays and thus expose the patient to lower energy. They are also done in the supine position which may reduce the chance of inter-scan variability due to patient pose. MRI scans expose patients to no harmful radiation at all and also allow the spinal curve to be measured in 3D, which may aid clinical decision making. They are however more expensive to perform than both DXAs and X-rays.

In this work we use a registration method proposed in [Windsor et al. 2021]. This method learns correspondences between different modality scans by learning to match together scans belonging to the same patient. While in this work and the original paper anonymised data from the UK Biobank is used, such a method could could have potential negative impacts for patient de-identification if used

maliciously; one could use it to match together subjects from different datasets which could affect anonymisation. However, for the models trained here this could only be done for Dixon MRIs and DXA scans.

This method shows promise in that it demonstrates that, in some cases, one can perform 3D segmentation of anatomical structures with only 2D annotations, thus significantly reducing annotation burden. However, the method remains in its infancy and should not be used for any clinical applications until it receives further evaluation.

6.5 Acknowledgements

Rhydian Windsor is supported by Cancer Research UK as part of the EPSRC CDT in Autonomous Intelligent Machines and Systems (EP/L015897/1). We are also grateful for support from a Royal Society Research Professorship and EPSRC Programme Grant Visual AI (EP/T028572/1).

Chapter 7

Vision-Language Modelling for Radiological Imaging and Reports in the Low Data Regime

This chapter presents a paper awarded an oral presentation at MIDL 2023 (Nashville).

Vision-Language Modelling for Radiological Imaging and Reports in the Low Data Regime

Rhydian Windsor[1] Amir Jamaludin[1]
Timor Kadir[1,2] Andrew Zisserman[1]

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford

[2]Plexalis

Abstract

This paper explores training medical vision-language models (VLMs) – where the visual and language inputs are embedded into a common space – with a particular focus on scenarios where training data is limited, as is often the case in clinical datasets. We explore several candidate methods to improve low-data performance, including: (i) adapting generic pre-trained models to novel image and text domains (i.e. medical imaging and reports) via unimodal self-supervision; (ii) using local (e.g. GLoRIA) & global (e.g. InfoNCE) contrastive loss functions as well as a combination of the two; (iii) extra supervision during VLM training, via: (a) image- and text-only self-supervision, and (b) creating additional positive image-text pairs for training through augmentation and nearest-neighbour search.

Using text-to-image retrieval as a benchmark, we evaluate the performance of these methods with variable sized training datasets of paired chest X-rays and radiological reports. Combined, they significantly improve retrieval compared to fine-tuning CLIP, roughly equivalent to training with 10× the data. A similar pattern is found in the downstream task classification of CXR-related conditions with our method outperforming CLIP and also BioVIL, a strong CXR VLM benchmark, in the zero-shot and linear probing settings. We conclude with a set of recommendations for researchers aiming to train vision-language models on other medical imaging modalities when training data is scarce. To facilitate further research, we will make

our code and models publicly available.

7.1 Introduction

Recently, there has been much progress in the field of vision-language modelling (VLM) where powerful, aligned, visual and language representations, such as CLIP and ALIGN, are learnt from image-caption pairs scraped from the internet [Radford et al. 2021; Jia et al. 2021; Ilharco et al. 2021]. This has exciting implications for deep learning in medical imaging, where annotating large datasets for supervised training has been a long-standing challenge since it generally requires an expert annotator whose time is expensive and limited. Fortunately, almost all scans taken in a clinical setting will have a corresponding radiological report, describing key findings in free-text. VLMs can use these reports as a supervisory signal to learn representations of images, forgoing the need for manual annotation. However, popular general-domain VLM methods use incredibly large datasets of paired images and text – CLIP, for example, is trained on 400 million image-text pairs. Clearly, this is not feasible in the medical domain, where dataset size is limited by factors such as scanner availability, concerns about patient privacy and the cost of obtaining images. For some types of imaging investigations, researchers can realistically hope for no more than a few thousand training image-text pairs, even if data is collected from multiple imaging centres.

To alleviate this problem, we investigate a set of methods to improve VLM performance for dual encoder models when the number of image-text pairs available for training is limited. We begin by proposing several candidate methods to achieve this (Section 7.2). We evaluate these methods by using them to train models on progressively smaller datasets of paired chest X-rays and reports (Section 7.3). The aim is to find methods with minimal degradation in performance when the training set size is shrunk, as measured by retrieval on paired image-text data unseen during training. We show that with appropriate pretraining and supervision, a dual-encoder trained on thousands of paired image-text samples can achieve a performance comparable to one trained on hundreds of thousands of paired samples. Having identified the best performing training methods for retrieval, we then explore whether the improved retrieval corresponds to better performance

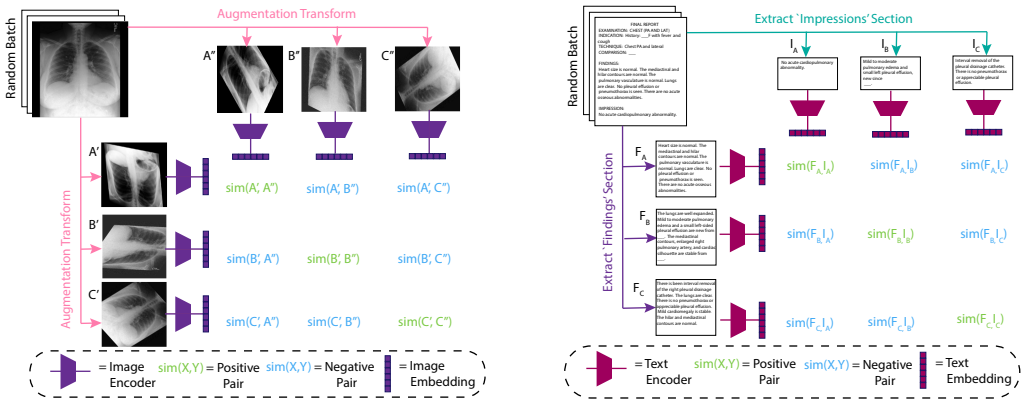
on downstream tasks, namely zero-shot classification of common conditions associated with chest X-rays. We conclude by offering a set of recommendations for researchers aiming to train VLMs on novel medical imaging domains with limited data (Section 7.4).

7.1.1 Related Work

In this work, we target dual encoder VLMs, that learn a joint visual-language representation, rather than generative models that ingest images and output text such as [Alayrac et al. 2022] for generic images and text or [You et al. 2021; Nooralahzadeh et al. 2021; S. Yang et al. 2022; Kayser et al. 2022] for automatic medical image report generation.

Several works have been published relating to learning joint representations of medical images and reports, mostly in the domain of chest X-rays (CXRs) – This is likely due to the existence of multiple large-scale, publicly available CXR datasets such as MIMIC-CXR [Johnson et al. 2019], CheXpert [Irvin et al. 2019] and PadChest [Bustos et al. 2020]. The majority of these works aim to learn representations of images and their associated reports using a contrastive learning paradigm (i.e. matching associated image-text pairs together across a randomly sampled batch) [X. Wang et al. 2018; Huang et al. 2021; Liao et al. 2021; Müller et al. 2022; Boecking et al. 2022; Zhang et al. 2022]. Several works have demonstrated that strong performance at this pre-training task correlates with performance across a wide range of downstream tasks, such as classification, segmentation, and natural language inference tasks. While many of these works explore performance when data for downstream tasks is limited, the setting of limited data for VLM training remains relatively neglected. This is a particularly important problem for medical imaging since few modalities other than CXR have publicly-available datasets of 100,000s image-text pairs. A few recent works [Segal et al. 2022; Y. Li et al. 2022; Mu et al. 2021] reduced bimodal training data for generic image-text pairs. However, these works consider subsets of exceptionally large datasets (of the order of millions of image-text pairs), rather than 1000s as considered here.

7.2 Increasing Data Efficiency During Vision-Language Modelling



(a) SimCLR Pretraining (b) CXR-BERT Specialization

Figure 7.1: The unimodal pretraining methods used to domain adapt image and text encoders before VLM training.

In this section, we outline possible methods for achieving strong dual encoder VLM performance when a limited number of *paired* image-text data are available for training. However, as is often the case, a large number of *unpaired* (unimodal) data may well be available, and we will take advantage of this to adapt the image and text encoders to the target domains before paired training begins. Broadly, the methods can be separated into two stages: (i) choosing an optimal dual encoder initialization by using image and text models trained to perform related unimodal self-supervised tasks; and (ii) adding more supervision during paired (bimodal) training by varying the contrastive loss function and also by generating additional positive image-text pairs. These two stages are described below:

Stage 1: Domain-adapting image and text encoders: In this stage, we start with strong, generic image and text encoders (e.g. an ImageNet-pretrained image model, and a BERT style language model pre-trained on large corpora). The encoders are then adapted to the CXR domain by self-supervised unimodal training as follows:

Image Encoder: To pretrain the image encoder, unimodal domain adaption is achieved using SimCLR [T. Chen et al. 2020]. In this method, multiple views of each image in the dataset are created using a range of augmentations. The image

encoder is then trained contrastively such that embedding vectors corresponding to different views of the same original image should have high cosine similarity. This process is illustrated in Figure 7.1a.

Text Encoder: To train a domain-adapted text encoder, we follow the suggestions of [Boecking et al. 2022]. We begin with CXR-BERT, a BERT model [Devlin et al. 2019] and tokenizer trained using standard RoBERTa-style masked language modelling (MLM) on two large corpora of generic clinical text datasets (PubMed Abstracts and MIMIC clinical notes), with a comparatively small amount of domain-specific data added in (MIMIC-CXR reports). The resulting model is then 'specialized' to chest X-ray reports by an additional pre-training step, whereby the language model is trained to contrastively match a report's 'findings' section to the 'impressions' section (i.e. summary) from the same report. Conceptually, this is analogous to SimCLR, except instead of different views of the same image, different sections from the same report (which should contain the same information) are matched together. This step is shown in Figure 7.1b.

Note that both these methods allow potentially larger datasets of unpaired images or reports to be used in training the model, alleviating the problem of limited paired data.

Stage 2a: Local vs. Global Loss Functions: Standard VLM training methods such as CLIP and ALIGN represent each image and each caption as a single global embedding vector and attempt to maximise the similarity between vectors from the same pair, usually via the InfoNCE loss function [Oord et al. 2018]:

$$\mathcal{L}_{InfoNCE} = -\frac{1}{2B}\sum_{i=0}^{B}\left[\log\frac{\exp(-\mathbf{v}_i.\mathbf{t}_i/\tau)}{\sum_j^B \exp(-\mathbf{v}_i.\mathbf{t}_j/\tau)} + \log\frac{\exp(-\mathbf{v}_i.\mathbf{t}_i/\tau)}{\sum_j^B \exp(-\mathbf{v}_j.\mathbf{t}_i/\tau)}\right]. \quad (7.1)$$

Here, $\mathbf{v}_i, \mathbf{t}_i \in \mathbb{R}^E$ represent E-dimensional L2-normalized embedding vectors of the image and text respectively for the i-th image-text pair in a batch of size B. τ represents the softmax temperature.

However, in medical imaging, reports often contain multiple unrelated statements, each referring to distinct regions within the image (e.g. 'left lung base atelectasis, mild cardiomegaly, right lung normal'). Since global embeddings are usually

produced by a pooling operation across the whole image, maintaining this spatial information in the embeddings is challenging. A possible solution to this problem are *local representations*, whereby image patches and text tokens are each represented by their own embedding vector. The model is trained such that the mean maximum patch-to-token and token-to-patch similarity is high for matching image-text pairs and low otherwise. In this paper, we consider GLoRIA, the local loss function proposed in [Huang et al. 2021]. Concretely, for a image-text pair with W word embeddings, $\mathbf{t} \in \mathbb{R}^{W \times E}$ and local embedding vectors describing M image regions, $\mathbf{v} \in \mathbb{R}^{M \times E}$ (again both L2-normalized), a similarity matrix $S \in \mathbb{R}^{W \times M}$ is calculated, where s_{ij} represents the cosine similarity between the i-th word embedding and j-th image region embedding. Using this, an attention score is calculated,

$$a_{ij} = \frac{\exp(s_{ij}/\tau_2)}{\sum_{k=1}^{M} \exp(s_{ik}/\tau_2)}. \tag{7.2}$$

This is then used to calculate an attention-weighted representation of the image for word i, $\mathbf{c}_i = \sum_{j=0}^{M} a_{ij}\mathbf{v}_{ij} \in \mathbb{R}^E$. The total report-image similarity is given by the matching function

$$Z(\mathbf{v}, \mathbf{t}) = \log(\sum_{i=1}^{W} \exp(\mathbf{c}_i.\mathbf{t}_i/\tau_3)^{\tau_3}. \tag{7.3}$$

$Z(\mathbf{v}, \mathbf{t})$ then replaces $\mathbf{v}.\mathbf{t}$ in equation 7.1, giving the following symmetric loss for a batch:

$$\mathcal{L}_{GLoRIA} = -\frac{1}{2B} \sum_{k=1}^{B} \left[\log \frac{\exp(-Z(v,t)/\tau)}{\sum_{l}^{B} \exp(-Z(v_k, t_l)/\tau)} + \log \frac{\exp(-Z(v,t)/\tau)}{\sum_{l}^{B} \exp(-Z(v_l, t_k)/\tau)} \right]. \tag{7.4}$$

As in the original GLoRIA paper, we also explore a simple combination of global and local loss functions, i.e. $\mathcal{L} = \mathcal{L}_{InfoNCE} + \mathcal{L}_{GLoRIA}$.

Stage 2b: Additional Supervision During VLM Training: As well as varying the form of the contrastive loss function, we investigate two further adaptions to VLM training. These are: (a) adding text-only & image-only self-supervision terms to the total loss; (b) creating more positive image-text pairs via augmentation and nearest-neighbour search. To do this, we use the Data efficient CLIP

(DeCLIP) framework [Y. Li et al. 2022]. The overall process is illustrated in Figure 7.2. For image-only self-supervision, SimCLR is used (as in Stage 1). Text-only self-supervision is provided by masked-language modelling. Additional positive text-image pairs are generated as follows: Starting with a matching image-text pair, $\{\mathbf{v}_i, \mathbf{t}_i\}$, random-crop augmentations are used to generate another view of the image, $\mathbf{v}'_i$. Similarly, another 'view' of the text, $\mathbf{t}'_i$, is created by EDA [Wei and Zou 2019] text augmentation, involving random synonym replacement, word insertion, deletion, and position swapping. Finally, a nearest neighbour to the text, $\mathbf{t}_i^{NN}$ is sampled from the data as follows: text embeddings are cached during training in a first-in, first-out (FIFO) queue. Once the queue is populated, the nearest neighbour for each text embedding is found by cosine similarity and used as a positive example for both corresponding images. This results in 6 positive pairs which can be used for contrastive training; $\{\mathbf{v}_i, \mathbf{t}_i\}, \{\mathbf{v}'_i, \mathbf{t}_i\}, \{\mathbf{v}_i, \mathbf{t}'_i\}, \{\mathbf{v}'_i, \mathbf{t}'_i\}, \left\{\mathbf{v}_i, \mathbf{t}_i^{NN}\right\}$ and $\left\{\mathbf{v}'_i, \mathbf{t}_i^{NN}\right\}$. This framework is agnostic to the choice of the contrastive loss function, and can be adapted for both local (e.g. GLoRIA) and global (e.g. InfoNCE) matching functions. For a more detailed explanation of this method and the hyper-parameters used, see Appendix 7.5.3.

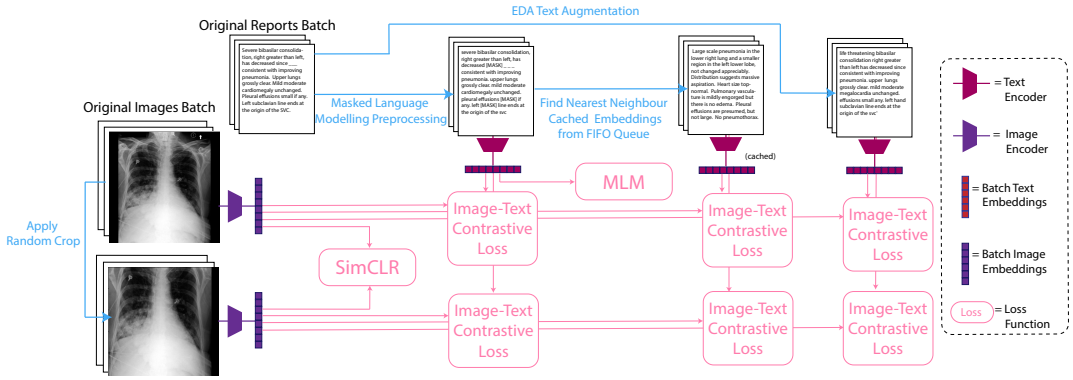


Figure 7.2: The DeCLIP framework used to add additional supervision during VLM training. Note that the image-text contrastive loss function can be either local or global.

7.3 Experiments

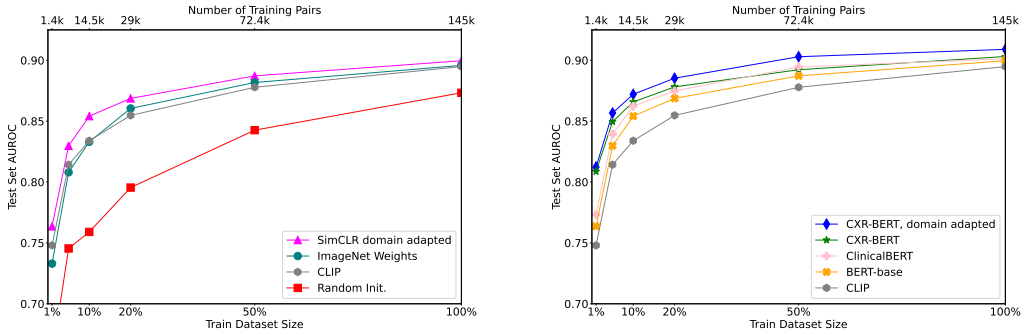
Dataset: For the experiments in this paper, we use the MIMIC-CXR-JPG dataset. [Johnson et al. 2019]. We include all reports with at least one frontal X-ray (AP or PA) and an 'Impression' section in the associated report, leaving 181,112 stud-

ies (144,781/18,679/17,652 in the training, validation, and test sets respectively). Each image is resized to 512×512 pixels by bilinear interpolation. For the reports, we used only the 'Impression' section as input to the text encoder during VLM training, however, the 'Findings' section is also used when domain-adapting the text encoder. We leave incorporating the longer 'Findings' sections into VLM training as future work. The details of the training, testing and validation splits used are given in Appendix 7.5.1.

Architecture, Evaluation metrics & Implementation details: In all experiments, we use a ResNet50 [He et al. 2016] image encoder and a BERT-style text encoder. All configurations are trained using an Adam optimizer with a learning rate of 10^{-3} and $\beta = (0.9, 0.99)$ with a batch size of 20 and softmax temperature of 0.5. Models are trained for 500 iterations and then evaluated for 100 iterations on the same subset of the validation set. Training continues until the validation loss does not improve for 10 successive validation epochs. Further details on each stage of training, including augmentations used, computational resources and training times are given in Appendix 7.5.2. We evaluate models by measuring text-to-image retrieval on the test set after training has concluded. This involves calculating the similarity between each image and report in the test split and measuring true and false positive rates at a range of similarity thresholds. We report the area under the resulting ROC curve (AUROC) for models trained with 1, 5, 10, 20, 50, and 100% of the possible training pairs.

Stage 1: Varying Dual Encoder Initialization: Figures 7.3a and 7.3b show results of experiments varying the weight initialization for both the image and text encoders, before training the dual-encoder using an InfoNCE contrastive loss. For the experiments varying image encoder pretrained weights, a standard uncased BERT base model is used to initialize the text encoder. We experiment with three image schemes: random initialization, standard ImageNet pretrained weights, and ImageNet weights with further domain adaption via image-only SimCLR pretraining. For the experiments varying the text encoder initialization, the SimCLR domain adapted image encoder is used. We compare four BERT models as text initializations: uncased BERT-base, ClinicalBERT, and CXR-BERT

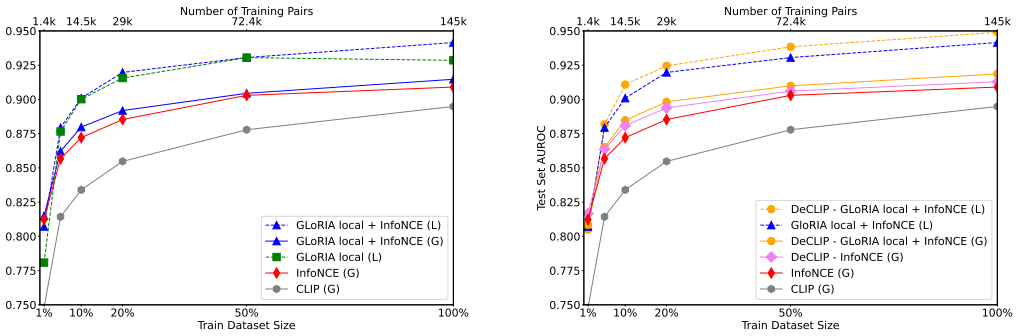
with and without further domain adaption. For both sets of experiments, we also plot the results of initializing the text and image encoders with ResNet50-CLIP pretrained weights. For both the images and text, initializing the encoder using weights from models trained via uni-modal self-supervised domain-adaption shows clear benefits. This is true at all dataset sizes, however, the effects are particularly pronounced with smaller training dataset sizes (1-20% of the data). For this reason, the SimCLR-domain adapted ResNet and domain-adapted CXR-BERT are used for initialization in all subsequent experiments unless otherwise stated.



(a) Varying Image Encoder Initialization (b) Varying Text Encoder Initialization

Figure 7.3: Training the dual encoder starting from various initializations for different size paired training data. (a) Varying the image encoder initializations, with a BERT-base-uncased as the initial text encoder. (b) Varying the text encoder initializations, with a simCLR-domain adapted ResNet50 model as the initial image encoder. InfoNCE contrastive loss is used to train the VLM, as in CLIP.

Stage 2a: Local vs. Global Loss Functions: Here, we explore varying the form of contrastive loss function used to match images and text. In total, three configurations are tested: (i) InfoNCE (equation 7.1); (ii) the local loss function component of GLoRIA (equation 7.4); and (iii) a combination of (i) and (ii). Note that for local loss functions, calculating pairwise similarities requires cross-attention between the image and text embedding vectors (see equations 7.2 and 7.3). This retrieval operation is generally more accurate however much slower and less suited to large scale datasets (see [Miech et al. 2021] for further discussion of this issue). In this paper, we report retrieval performance using both this local, cross attention mechanism (*local matching*, marked '`(L)`' in Figure 7.4) as well as via global embedding vector cosine similarity (*global matching*, marked '`(G)`'). For methods that have both global and local objectives, both matching functions are reported.



(a) Local vs. Global Loss Functions (b) DeCLIP vs. standard VLM training

Figure 7.4: Varying the contrastive loss function used during VLM training. (a) Evaluates various local and global loss functions (b) DeCLIP evaluated against the standard InfoNCE approach. Methods performing retrieval using a local matching function `(L)` are shown as dashed lines while retrieval using a global matching function `(G)` are shown as solid lines. In all cases, the simCLR domain-adapted ResNet50 and domain-adapted CXR-BERT from stage 1 are used for initialization.

Figure 7.4a shows a comparison of the three configurations. As before, we plot the performance of CLIP initialization with an InfoNCE objective for comparison purposes. Our results show local matching results in better retrieval performance than global matching at all except very small training dataset sizes (1%). Crucially, we find that the extra supervision provided by both a local and global component to the contrastive loss function (`GLoRIA local + InfoNCE`), improves retrieval at both global and local matching, compared to respective methods which use a local (`GLoRIA local`) or global (`InfoNCE`) loss alone.

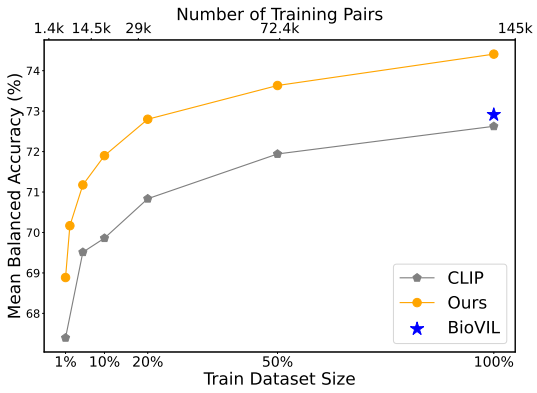
Stage 2b: Additional Supervision via DeCLIP: Here, we investigate the effect of adding more supervision via the DeCLIP framework, shown in Figure 7.2. This is done using two forms of contrastive loss function: (a) InfoNCE alone; (b) InfoNCE and the local component of GLoRIA. The trained models are compared to those trained by equivalent methods, but without the extra supervision given by DeCLIP. The results of this experiment are shown in Figure 7.4b. Encouragingly, DeCLIP improves retrieval performance in all settings at almost all dataset sizes. This is also additive with the improvement found by using a global and local loss function shown in Figure 7.4a. An ablation study of each component of DeCLIP is shown in Appendix 7.5.3, along with example nearest-neighbour reports.

Overall, with the benefits of self-supervised pretraining (Stage 1) and the combination of global and local loss functions with additional DeCLIP supervision

(Stage 2), the retrieval curve rises quickly in the low-data regime, approaching a horizontal asymptote. For example, for 10% of the training data, the performance achieves over 95% of the final value (using 100% of the data). Also, the combination of methods far exceeds the fine-tuned CLIP baseline. Example retrieval results for this model are shown in Appendix 7.5.2.

Evaluation on Downstream Tasks: The experiments above suggest that several methods can be employed to improve global text-to-image retrieval on the test dataset. Here, we explore how these improvements correlate to downstream classification tasks. Specifically, classifying 12 common CXR-related conditions using labels from the CheXpert labeller [Irvin et al. 2019]. The full details of this experimental set-up, as well as class-level results, are given in Appendix 7.5.5. We evaluate our best retrieval model (`DeCLIP - InfoNCE + GloRIA local`) at zero-shot image classification and also under linear probing, with the results shown in Figure 7.5. For comparison, we report the performance of the CLIP-initialized models trained using InfoNCE alone and also BioVIL [Boecking et al. 2022], a strong, publicly-available VLM for CXRs which is also trained on MIMIC-CXR[1]. From Figure 7.5, one can see that the improvements in retrieval carry over to the downstream evaluation, with our method achieving better performance than CLIP models trained on 5-10× the data in the linear probing setting (e.g. ours trained with 1% achieves 70.1% accuracy compared to 69.9% for CLIP trained on 10%) The zero-shot results are noisier, however again our method improves on CLIP and BioVIL at all dataset sizes.

[1]The original paper does not report a train-test split, thus BioVIL may be trained on some image-text pairs which appear in our test set.



Dataset Size		% Average Balanced Accuracy			
		Zero-Shot		Linear Probe	
Frac.	# Pairs	CLIP	Ours	CLIP	Ours
1%	1,448	60.8	65.7	67.4	70.1
5%	7,239	64.4	66.2	69.5	71.2
10%	14,478	64.6	66.4	69.9	71.9
20%	28,956	66.7	66.7	70.8	72.8
50%	72,390	65.7	68.7	71.9	73.7
100%	144,781	65.3	68.5	72.6	74.5

Figure 7.5: Zero-shot classification and linear probing of common conditions using CheXpert labels on the test split. BioVIL achieves balanced accuracy scores of 67.3% and 72.9% in the zero-shot and linear probe settings respectively. Note that 'CLIP' here refers to CLIP-initialization finetuned on the same dataset using and InfoNCE loss alone.

7.4 Conclusion

This paper has explored several methods of improving VLM-models when image-text pairs are limited, a particularly relevant problem to medical imaging. Based on these results, we make the following recommendations: (a) unimodal, self-supervised 'domain-adaption' of generic image and text models before VLM training vastly improves retrieval performance at all dataset sizes; (b) a combination of global and local contrastive loss functions is also beneficial; (c) adding more supervision during VLM training, via uni-modal self-supervision and generated additional positive image-text pairs, is another method to increase performance. Combining these ideas, we achieve state-of-the-art performance at zero-shot classification on MIMIC-CXR-JPG when trained on the whole dataset, and maintain strong performance even with limited training data. We hope these results will be useful to researchers aiming to train VLMs on medical imaging domains, especially when training data is scarce. To aid further research in this area, our code and models will be made publicly available.

7.5 Appendix

7.5.1 Dataset Details

In this paper, we used the MIMIC-CXR-JPG dataset for training and testing. This is a large dataset of chest X-rays and associated radiological reports. DICOMs from the original MIMIC-CXR dataset are converted into JPG files[2] and the CheXpert labeller [Irvin et al. 2019] is used to extract labels from the reports. The dataset is split into training, validation and testing splits on the patient-level, as shown in Table 7.6.

Dataset Split	# Unique Subjects	# Report-Scan Pairs	Folders
Total	65,398	180,650	p10-p19
Train	52,428	144,319	p12-p19
Validation	6,572	18,679	p11
Test	6,398	17,652	p10

Figure 7.6: The splits used for MIMIC-CXR-JPG in the experiments in this paper.

7.5.2 Additional Training Details

Augmentation Hyperparameters

Table 7.1 reports augmentation hyperparameters for all training stages used in this paper. We report the augmentations for both the image domain-adaption stage using SimCLR and for full VLM training. Note that the DeCLIP framework introduces an additional image augmentations by producing an augmented version of the original image for a simCLR-like unimodal supervision objective. In this case we simply use a 50% random crop of the already-augmented original image, maintaining the same aspect ratio.

Training Time & Computational Resources

Most models are trained with a batch size of 20 using 2× 24GB NVIDIA Tesla P40 GPUs. VLM training using the DeCLIP framework requires an additional GPU

[2]Further explanation of the MIMIC-CXR-JPG dataset is given at https://physionet.org/content/mimic-cxr-jpg/2.0.0/

	VLM Training	SimCLR Image Pretraining
Translation	±5px	±20
Rotation	−10 to 10°	−180 to 180°
Brightness	±20%	±20%
Shear (x-axis)	-	40°
Scaling	±10%	±10%
Horizontal Flip	-	$p = 0.5$
Gaussian Noise	±5%, $p = 0.5$	±5%, $p = 0.5$
Gaussian Blur	-	$\sigma = (1, 3, 5)$, $p = 0.5$

Table 7.1: Image augmentation parameters used during VLM training and while domain adapting the image encoder using simCLR.

of the same description to maintain the same batch size. Approximate training times for each stage are given in Table 7.2.

Training Stage	Additional Information	# Samples	Training Time (hours)
Image Encoder Domain Adaption	-	144,319	12
Text Encoder Domain Adaption	-	144,319	4
VLM Training (100%)	CLIP framework, InfoNCE loss	144,319	9
VLM Training (100%)	CLIP framework, Combined loss	144,319	10
VLM Training (100%)	DeCLIP, InfoNCE loss	144,319	20
VLM Training (1%)	DeCLIP, Combined loss	1,443	2.5
VLM Training (10%)	DeCLIP, Combined loss	14,432	9
VLM Training (100%)	DeCLIP, Combined loss	144,319	36

Table 7.2: Approximate training times for each stage of our training pipeline. Note that for the image and text encoder domain adaption stages, the models are initialized using weights from generic pretraining (ImageNet & CXR-Bert respectively).

7.5.3 DeCLIP Method, Nearest Neighbours & Ablation Study

Data Efficient CLIP (DeCLIP) [Y. Li et al. 2022] is a method for training CLIP-like models aimed at maintaining strong performance with less image-text pairs. The main idea behind this method is to add several additional forms of supervision during VLM training, alongside the standard image-text matching contrastive InfoNCE loss. This additional supervision can be broadly seperated into three classes; uni-modal self-supervision, multi-view supervision and nearest-neighbour supervision. These additional loss terms can be seen in Figure 7.2 and are described individually below:

Uni-Modal Self-Supervision: One can add in any form of image-only or text-only self-supervision in addition to the contrastive image-text matching function. We opt for two common image and text objectives, namely SimCLR [T. Chen et al. 2020] using random crop augmentations and masked language modelling, normally used to train BERT-like language models from scratch. This also acts as a form of text augmentation during VLM training since each token is randomly masked out or replaced with another random tokens with some small probability, p.

Multi-View Supervision: This component of the framework aims to add additional supervision by creating additional 'views' of the data; augmentations of the images, I', and text, T', with different appearance but identical semantic meaning. These can then be used to create three new image-text pairs for VLM training: $(I, T'), (I', T)$ and (I', T'). I' is generated using the same random-crop augmentation used for simCLR image self-supervision described above. T' is generated using the EDA method proposed by [Wei and Zou 2019]. This is a sentence-level simple text augmentation method which randomly: (a) selects n_{syn} words from the text and replaces them with synonyms; (b) selects n_{ins} words and inserts synonyms of them at random positions; (c) randomly swaps the positions of n_{swap} word pairs (d) randomly deletes words with p_{del}. For a sentence of W words, we use $n_{syn}, n_{ins}, n_{swap} = 0.1 \times W$ and $p_{del} = 0.1$. This is done using the official implementation of the method[3].

Nearest-Neighbour Text Supervision: Augmentation provides one method of generating new semantically-similar alternatives to the text and images. Another approach is to sample other images or reports from the training dataset which are projected into similar points in latent space by the dual encoder. We do this for text reports from the dataset. Practically, this is done by storing the last N text embeddings produced by the text encoder in a first-in, first-out (FIFO) queue. Then for each report embedded, the nearest neighbour, $\mathbf{t}_{NN}$, can be found by measuring the maximum similarity between the report's global embedding vector and all elements in the queue. This can then be used to create two new positive pairs

[3] https://github.com/jasonwei20/eda_nlp

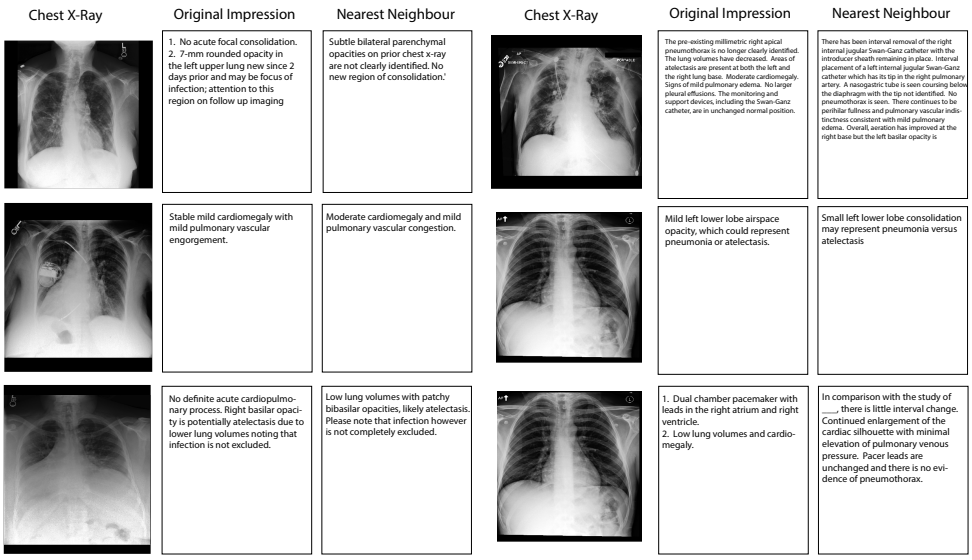


Figure 7.7: Example report nearest-neighbours found via the DeCLIP method using the model trained on 100% of the data. The text encoder finds nearest neighbour reports with the same semantic information as the original impressions, said in different words.

for VLM training, using the original and random-cropped images. Some examples of nearest neighbours found via this method are shown in Figure 7.7.

Combined Loss Function: The result is three additional components to the overall VLM loss function - the losses from text and image self-supervision, $\mathcal{L}_{TSS}$, $\mathcal{L}_{ISS}$, the mean loss from the three additional image-text pairs introduced using multi-view supervision, $\mathcal{L}_{MVS}$ and from the two additional pairs generated by nearest neighbour supervision, $\mathcal{L}_{NN}$. These are combined as follows:

$$\mathcal{L} = (1 - \alpha - \beta - \gamma)\mathcal{L}_{orig.} + \frac{\alpha}{2}(\mathcal{L}_{TSS} + \mathcal{L}_{ISS}) + \beta\mathcal{L}_{MVS} + \gamma\mathcal{L}_{NN}. \tag{7.5}$$

Here, $\mathcal{L}_{orig.}$ is the contrastive loss function for the original image-text pair. This can be either InfoNCE, the local component of GLoRIA, or a combination of the two. In the experiments in this paper we use $(\alpha, \beta, \gamma) = 0.2$.

Ablation Study: Figure 7.8 shows an ablation study where each of these components are studied in isolation, by setting two of α, β and γ to zero and the other to 0.5, corresponding to the component being tested.

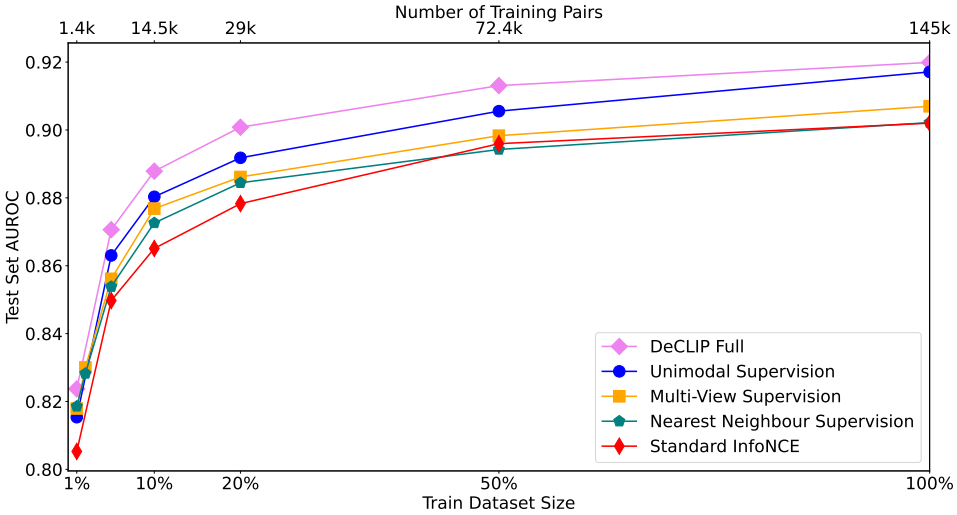
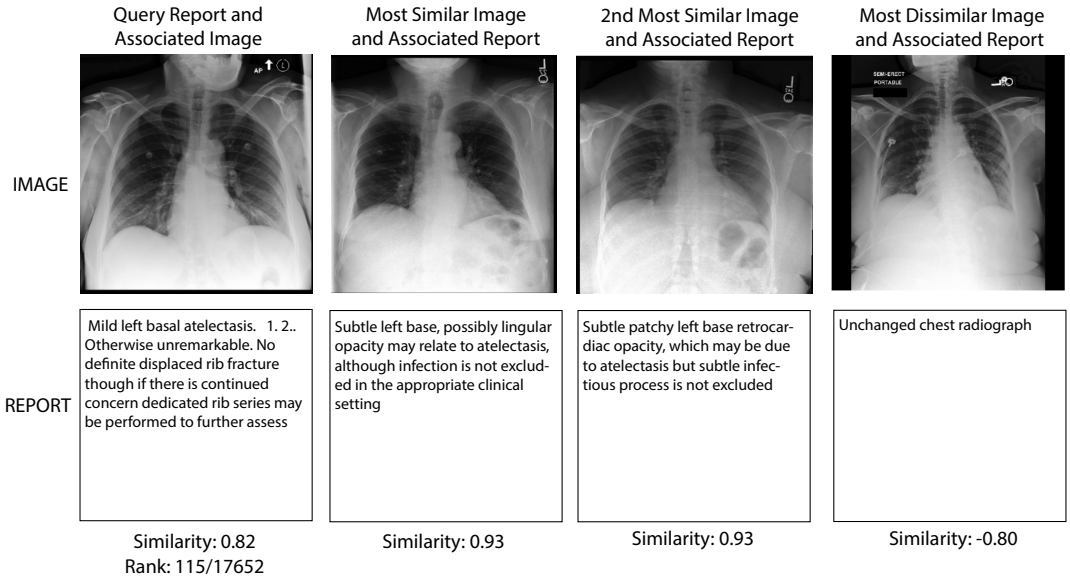


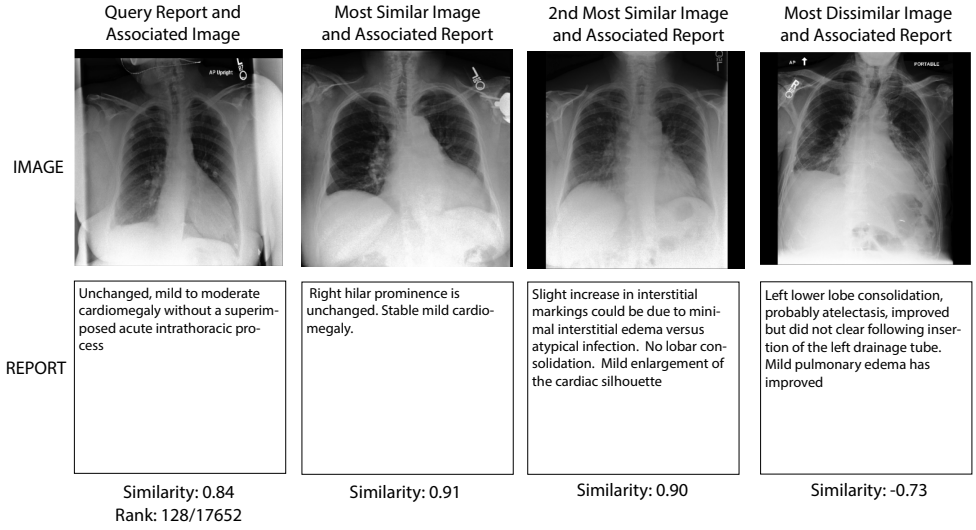
Figure 7.8: Ablation study for each individual component of the DeCLIP framework. In each case, depending on which component is being studied, two of (α, β, γ) from Equation 7.5 are set to 0, while the other equals 0.5.

7.5.4 Qualitative Retrieval Results

Figure 7.9 shows text-to-image retrieval examples for the best performing model, using the DeCLIP framework with a combined global and local loss. Each report is embedded and the global similarity with all images in the test dataset is calculated. Along with the query report and its associated image, we show the the retrievals at $k = 1, 2$, as well as the least similar image. From these results one can see that the trained model performs recall based on key conditions mentioned in the report, with almost all of the matching images having the same conditions present as those mentioned in the query.

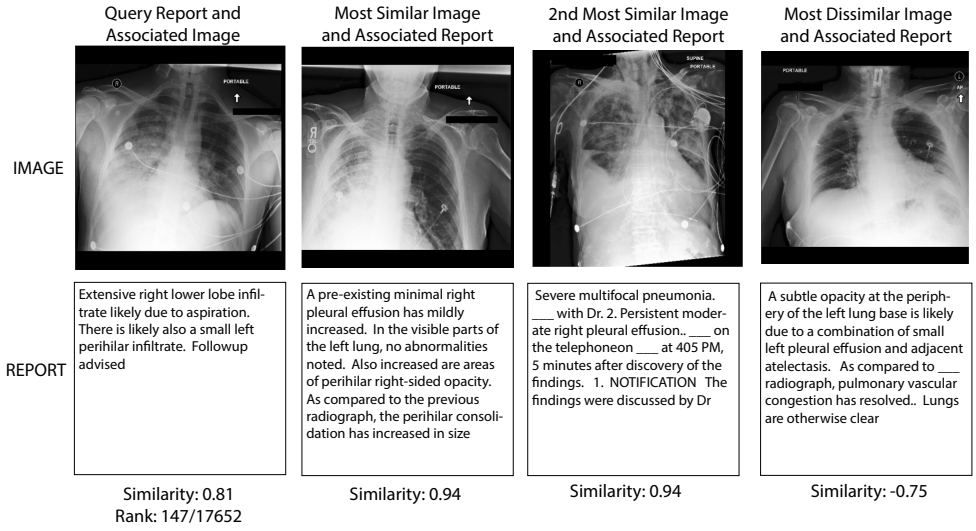


(a) Example query report demonstrating atelectasis. Note that first and second most similar images also exhibit atelectasis in the left lung base (as mentioned in the corresponding reports).

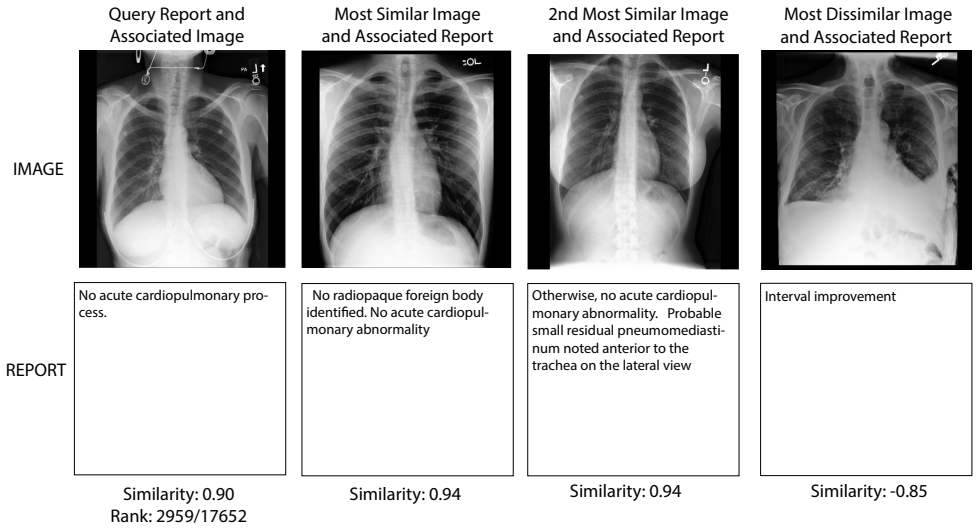


(b) Example report with mild cardiomegaly (enlarged heart). Again, the retrieved examples also exhibit mild cardiomegaly.

Figure 7.9: Example text-to-image retrieval results for the final model. All images in the test dataset are ranked based on similarity to the query report. Based on this criteria, the first, second and least most similar image-text pairs are shown, along with the original report and it's associated image.

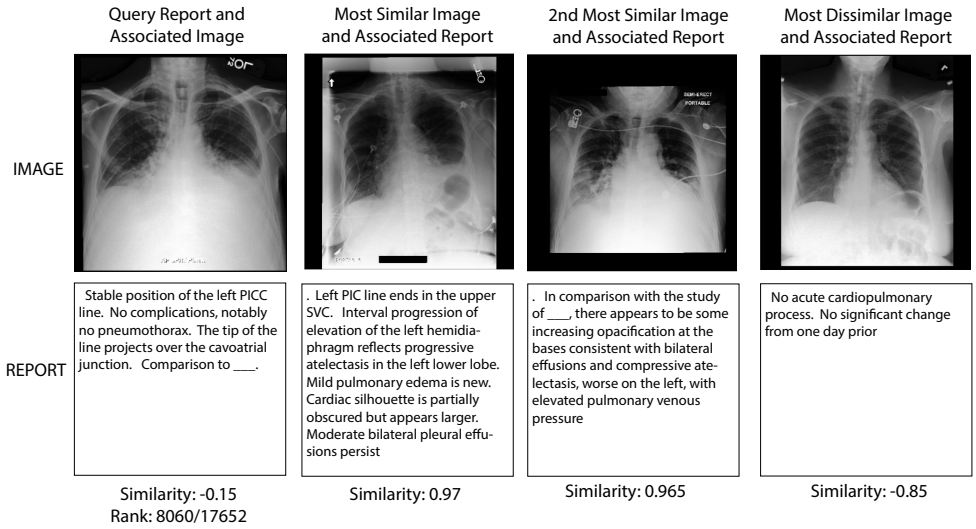


(c) Example query report with infiltration of the right lung.



(d) An example query reporting no findings. Since this is true of a large subset of the dataset, the retrieval rank is fairly low here ($2{,}959^{\text{th}}$ out of 17,652 candidate images), despite high similarity.

Figure 7.9: (continued) Example text-to-image retrieval results for the final model. All images in the test dataset are ranked based on similarity to the query report. Based on this criteria, the first, second and least most similar image-text pairs are shown, along with the original report and it's associated image.



(e) Example failure case. The original image has a very low similarity, perhaps because the PICC line is very challenging to see in this case.

Figure 7.9: (continued) Example text-to-image retrieval results for the final model. All images in the test dataset are ranked based on similarity to the query report. Based on this criteria, the first, second and least most similar image-text pairs are shown, along with the original report and it's associated image.

7.5.5 Downstream Task Evaluation Details

Zero-shot Classification and Linear Probing

We evaluate our models at zero-shot classification on our test-split of MIMIC-CXR. Specifically we attempt to classify 12 conditions labelled by the CheXpert labeller [Irvin et al. 2019]: Cardiomegaly, Atelectasis, Lung Opacity, Pleural Effusion, Edema, Pneumonia, Pneumothorax, Consolidation, Fracture, Enlarged Cardiomediastinum, Support Devices, Lung Lesion. We leave out the 'Pleural Other' and 'No Finding' classes, since these are difficult to write an all-encompassing prompt for - 'No Finding' since it indicates the absence of all conditions, rather than the presence of a specific one and 'Pleural Other' since it is very rare and can refer to a wide range of pleural disorders. Note that we treat all 'uncertain' cases, i.e. where the condition is not mentioned in the report, as assumed negatives.

Zero-shot Classification Method: For each class X, we embed two phrases using the trained text encoder; "X remains visible" and "There is no evidence of X", giving prompt embedding vectors $\mathbf{p}_X^+$ and $\mathbf{p}_X^-$. Then for a given report-image pair, the cosine similarity between the (normalized) prompt embeddings and image

global embedding vectors is measured, $\left\{\mathbf{v}_i.\mathbf{p}_X^+, \mathbf{v}_i.\mathbf{p}_X^-\right\}$. The zero-shot prediction for each class is then measured by softmax($\mathbf{v}_i.\mathbf{p}_X^+, \mathbf{v}_i.\mathbf{p}_X^-$).

Linear Probing Method: As an additional method to measure the quality of the image encoder's features, we perform linear probing on the test dataset. This is done by 5-fold cross validation, performing class-weighted logistic regression (since the dataset is very imbalanced) on 80% of the test data and testing on the remaining 20% of data. We report the mean balanced accuracy across all classes, averaged over all 5 folds.

The classification of both these methods is shown at the label-level in Table 7.3.

Training Pairs (%)	Model Name	Balanced Accuracy (Zero-shot/Linear Probe, %)							
		Cardiomegaly	Atelectasis	Lung Opacity	Pleural Effusion	Edema	Pneumonia	Pneumothorax	Consolidation
1	CLIP	45.1/68.7	61.5/68.2	58.6/70.2	75.8/78.3	56.0/61.1	56.7/54.0	54.4/61.4	56.2/65.1
	Ours	68.0/70.6	67.9/69.2	70.9/71.9	76.5/80.7	61.7/61.3	51.0/55.1	55.3/60.6	65.4/66.4
5	CLIP	65.9/69.7	65.3/68.5	70.2/72.1	74.2/79.7	59.6/63.3	54.1/59.2	55.7/64.0	62.1/66.1
	Ours	70.1/72.9	69.4/71.9	71.8/74.2	76.7/81.8	62.5/63.1	51.5/57.3	56.9/64.3	66.2/67.4
10	CLIP	67.8/70.5	66.7/70.5	66.7/72.3	77.1/80.0	56.6/62.9	53.4/58.6	57.5/63.5	60.9/66.2
	Ours	69.2/72.8	69.1/73.3	72.0/74.9	76.6/81.7	62.4/62.9	52.8/57.4	59.5/66.0	66.7/68.6
20	CLIP	68.1/72.2	68.4/70.4	70.5/72.8	73.7/80.7	62.4/65.5	53.2/59.2	58.7/63.1	66.1/66.7
	Ours	70.2/73.6	68.7/73.3	72.9/75.7	76.4/82.2	63.8/65.4	54.4/59.0	60.8/66.8	66.6/68.8
50	CLIP	69.3/72.9	65.0/72.7	70.8/75.1	71.6/81.3	59.6/65.6	50.5/59.9	58.6/64.9	66.6/68.0
	Ours	70.1/74.1	69.6/74.3	73.6/76.3	76.8/82.5	63.8/64.4	57.0/61.8	63.8/66.7	67.5/69.4
100	CLIP	69.9/73.7	68.2/73.4	71.8/75.5	74.7/81.3	63.9/65.2	54.6/60.5	49.8/67.6	66.7/68.1
	BioVIL	70.2/73.8	69.6/73.2	74.4/74.7	72.5/80.4	63.0/65.2	52.1/60.9	60.1/65.2	66.2/68.7
	Ours	70.1/74.7	69.5/74.6	74.1/76.5	75.0/83.0	63.3/66.1	57.3/60.7	63.7/69.2	67.2/70.0

Training Pairs (%)	Model Name	Balanced Accuracy (Zero-shot/Linear Probe, %)				
		Fracture	Enlarged Cardiomediastinum	Support Devices	Lung Lesion	**Mean**
1	CLIP	75.1/76.4	54.1/59.7	61.3/68.0	74.7/77.7	60.8/ 67.4
	Ours	76.7/79.8	59.9/63.5	64.3/69.8	70.5/77.8	65.7/ 68.9
5	CLIP	77.5/79.3	57.5/61.5	62.3/71.5	68.0/79.3	64.4/ 69.5
	Ours	79.1/81.4	60.5/65.0	62.7/74.3	67.1/80.3	66.2/ 71.2
10	CLIP	78.1/80.3	58.9/61.5	62.2/72.1	69.6/79.9	64.6/ 69.9
	Ours	79.2/82.1	61.6/66.1	61.9/75.4	66.4/81.6	66.5/ 71.9
20	CLIP	80.4/81.1	60.4/63.3	68.4/74.2	70.4/80.8	66.7/ 70.8
	Ours	79.0/82.6	62.1/66.5	57.7/76.8	68.2/82.8	66.7/ 72.8
50	CLIP	79.4/81.7	61.4/63.4	63.3/76.4	72.4/81.5	65.7/ 71.9
	Ours	80.8/83.2	62.4/67.0	68.0/79.4	70.7/84.4	68.7/ 73.6
100	CLIP	78.4/82.1	62.2/64.5	64.4/76.9	58.7/82.7	65.3/ 72.6
	BioVIL	79.4/82.4	62.0/66.1	70.1/78.7	68.6/83.9	67.3/ 72.9
	Ours	79.9/83.7	62.9/68.3	68.8/81.0	70.0/85.6	68.5/ 74.5

Table 7.3: Full class level results for the zero-shot and linear probe classification experiments using CheXpert labels on the test set. Balanced accuracy scores are reported as 'Zero-shot accuracy/Linear Probe accuracy' for each task.

Chapter 8

Discussion

In this chapter, we begin by summarising the impact the presented work has had so far on the research community. We then turn our attention to possible future directions for this work, specifically in the context of recent advancements in machine learning and computer vision.

8.1 Impact

Chapters 2 & 3 present novel methods to detect and label vertebrae in clinical spinal MRI scans. The work in Chapter 3 has received several citations from groups suggesting other architectures for vertebrae detection in MRI and CT [Zhao et al. 2022; Azad et al. 2021; F. Wang et al. 2021; Y. Chen et al. 2023], as well as attempting automated vertebral fracture detection [Zakharov et al. 2023] and scoliosis monitoring [Tapp et al. 2021]. Furthermore, the architecture proposed in Chapter 4 has also been adopted as a baseline for other works exploring architectures for automated radiological grading [Kowlagi et al. 2022].

Likewise, the work in Part II has received several citations from groups working on learning representations for medical imaging across multiple imaging modalities and tasks, for example in registration of retinal imaging [An et al. 2022], organ segmentation in CT scans [J. Li et al. 2022] and tumour classification in CT and Brain MRI scans [S. Chen et al. 2022]. Interestingly, it has also received attention outside medical imaging, with one study adopting the self-supervised learning framework presented in Chapter 5 for satellite imagery [Scheibenreif et al. 2022]

and another presenting a related method for radio sensing [Alloulah and Arnold 2023]. Finally, while the work presented in Chapter 7 has only just been accepted, however we hope it is of interest to a wide audience of researchers working on vision-language models in medical imaging.

SpineNet During my DPhil, I combined presented methods for vertebra detection and radiological grading into a single source-available software package called SpineNet[1]. This takes a T2 lumbar spinal MRI scan as input and outputs radiological grading scores for a range of common spinal conditions. This is the second iteration of SpineNet, with multiple improvements over the first version, including a far more accurate and faster vertebra detection and radiological grading pipeline (the entire process takes less than 2 seconds as opposed to a few minutes in the first version). Furthermore several new gradings are added. Perhaps most importantly, the software is now source-available. To the best of my knowledge, this is the first such tool to be publicly available to researchers for the task of automated radiological grading in clinical spinal MRI scans.

SpineNet been adopted by several researchers in the back pain community and one recent study [McSweeney et al. 2022] benchmarked it against experts on an external dataset of lumbar MRIs, concluding that SpineNet 'matched human reliability and demonstrates robust performance despite the multiple challenges facing model generalizability.' The software has also been used in industrial research projects, with a major pharmaceutical company adapting it to perform automated progression monitoring of Ankylosing Spondylitis (a severe type of arthritis) in clinical drug trials ([Jamaludin et al. 2020b] presents some early results of this on-going work). We have written a technical paper describing the SpineNet pipeline at the initial release and a journal paper describing the latest version of SpineNet is currently under review for Scientific Reports.

[1]Web Demo: http://zeus.robots.ox.ac.uk/spinenet/index.html, Source Code: https://github.com/rwindsor1/SpineNet

8.2 Future Work

Over the course of the last four years, the research landscape in machine learning has changed signficantly and this affords several exciting opportunities to build on and extend the work presented in this thesis.

8.2.1 Using Free-Text Radiological Reports To Learn Representations of Spinal MRIs

In this thesis we have explored radiological grading of clinical spinal MRI scans and also the use of radiological reports as a form of weak supervision for learning representations in medical imaging. A natural next step would be to combine these contributions, and use associated free-text radiological reports describing spinal MRI scans investigations as a form of supervision to train models to operate on these images. As previously discussed, this confers several advantages over tabular manual annotation, allowing for larger training datasets and a more flexible description of the scans.

One could imagine multiple ways of doing this. Firstly, recent advances in natural language processing (e.g. GPT-3 [Brown et al. 2020]) have yielded models proficient at converting unstructured text into a structured format, such as JSON or CSV files. Such models could be leveraged to convert free-text radiological reports into a structured format, producing an annotation table akin to those produced by typical manual annotation. An example of a preliminary attempt to perform this task is shown in Figure 8.1.

The major advantage of this approach is that it is considerably less labour intensive than manual annotation. However, there are several challenges to be overcome. Firstly, the quality of the conversion will depend on the quality of the language model used. While models such GPT-3 and GPT-4 [OpenAI 2023] are impressive across many tasks, they are trained mostly on generic non-technical text and hence may struggle with the highly specialised, idiosyncratic language used in radiological reports. Furthermore, reports often often express uncertainty or ambiguity, e.g. "lesions may represent metastases" or "there is a possibility of a fracture". Previous studies have shown that LLMs often struggle with such ambiguity and are prone

```
PROMPT:

For the following clinical report, output a structured JSON describing the findings.
Each key should be a lumbar vertebral level (L5/S1, L4/L5, L3/L4, L2/L3, L1/L2, T12/L1).
The key values should be a list of conditions from the following ONLY:
["Disc Bulging", "Nerve Compression", "Disc Herniation", "Endplate Defect"]
Do not add additional conditions to the list.
Output 'Not mentioned' if that level is not mentioned.

Please confirm you understand the instructions by summarising them.

LLM:
Yes, I understand the instructions. I need to parse a clinical report and
create a structured JSON output where each key represents a lumbar
vertebral level and the value is a list of conditions from a given
list of conditions. If a condition is not mentioned in the report,
the value for that key should be 'Not mentioned'.

PROMPT:
MRI Spine Lumbar and Sacral :
Technical factors: T1 and T2 sagittals and axials.

Findings:

Multilevel disk dehydration wit multilevel disk bulges and facet joint change.
At the L2-3 and L3-4 levels, there are small diffuse disk bulges and moderate
facet joint changes.  Some foraminal encroachment, but no nerve root compression.
At the L4-L5 level, the disk is burst and moderate
facet joint changes.  Posterior displacement and potential impingement
of the right L4 nerve root.  The L5 nerve roots traverse normally.
At the L5-S1 level, there is a focal right subarticular disk herniation
superimposed on a diffuse bulge.  Impingement on the right S1 nerve root is present.
The L5 nerve roots exit freely.
The top of the S1 vertebra is damaged.

The distal thoracic cord is normal in signal and calibre.
The conus and cauda equina are normal.

LLM:
{
"L5/S1": ["Disc Bulging", "Disc Herniation", "Nerve Compression"],
"L4/L5": ["Disc Herniation", "Nerve Compression", "Facet Joint Changes"],
"L3/L4": ["Disc Bulging", "Facet Joint Changes"],
"L2/L3": ["Disc Bulging", "Facet Joint Changes"],
"L1/L2": "Not mentioned",
"T12/L1": "Not mentioned"
}
```

Figure 8.1: An example of a radiological report converted into a structured annotation table by a large language model (ChatGPT-3.5). The model has correctly identified most relevant clinical information from the report, and correctly disambiguated "the disk is burst" as a disc herniation. However it has also added "Facet Joint Changes" despite it not being mentioned in the list of classes. It should also be noted that the L4/L5 nerve compression is ambiguous in this report, which this annotation method is not able to capture.

to hallucinating additional details [Ji et al. 2023]. These factors could introduce label noise into training signals and accordingly developing robust methods to avoid such errors is an important area of future work.

Another approach would be to train a bi-modal vision-language model to learn directly from the raw text of the radiological reports as demonstrated on chest X-ray in chapter 7. This approach not only resolves many of the issues above, but also another key issue of tabular annotations - that of selecting a fixed set of classes to annotate. There is a long tail of spinal conditions seen in clinical practice, often with complicated and overlapping symptomologies and thus representing these using a fixed set of categories is necessarily a simplification. Natural language on the other hand is far more flexible and better at expressing nuance in a particular study. There are two immediate methods that one could use to train such a model: (1) matching together an embedding of the text with an embedding of the scan image as in CLIP [Radford et al. 2021]; (2) framing the problem as an image captioning task and aiming to generate a radiological report from the scan itself. In both approaches, the main challenge would be adapting such methods to datasets of paired clinical MRI scans and associated radiological reports, which are likely to be much smaller than the datasets used in previous work on vision-language models. Hopefully, the work presented in Chapter 7 will provide a useful starting point for future research in this area.

8.2.2 Automated Prognosis For Spinal Cancers & Primary Site Prediction

Another interesting direction is to expand on the work presented in Chapter 4 on automated detection and monitoring of spinal cancer and related conditions (such as cord compression and vertebral fractures). While the relatively small, single-centre dataset used in this work was sufficient to produce promising initial results as outlined in the paper, larger datasets are required to train more accurate models. Furthermore, with additional training data, several ambitious extensions to this work are possible. In this section, we briefly outline two such follow-ups and a dataset we have acquired suitable for these projects.

The first objective is to predict the primary site of cancer by the pattern of metas-

1. **What is the vertebra type?**
   - Junctional (C2, C7-T2, T11-L1, L5-S1) **+3**
   - Mobile Spine(C3-C6, L2-L4) **+2**
   - Semi-Rigid (T3-T10) **+1**
   - Rigid (S2-S5) **0**
2. **Does the patient experience pain?**
   - Yes **+3**
   - Occasional **+1**
   - None **0**
3. **What is the type of lesion?**
   - Lytic **+2**
   - Mixed **+1**
   - Blastic **0**
4. **How is spinal alignment?**
   - Subluxation/Translation Present **+4**
   - Hyperkyphosis/Scoliosis **+2**
   - Normal **0**
5. **How much of the vertebrae is involved? Is it collapsed?**
   - >50%, collapse **+3**
   - <50%, collapse **+2**
   - No collapse, >50% involved **+1**
   - No collapse, <50% involved **0**
6. **Is there posterior involvement?**
   - Bilateral **+3**
   - Unilateral **+1**
   - None **0**

Table 8.1: The Spinal Instability Neoplastic Score (SINS) [Fisher et al. 2010], a heuristic scoring method to determine whether a tumour-infiltrated vertebra should be stabilised through surgical intervention. Each question is answered for the metastasis of interest and the scores are summed. Surgical intervention is usually recommended for scores of 7 or more.

tases produced. In some cases, the first indication of cancer is metastases in the spine and the primary site is unknown. Clearly, this is a very important piece of information for clinical decision making as different primaries require different treatments. If this could be done through imaging, it reduce the need for biopsy-based tests which are invasive, expensive [Kerr et al. 2012] and often inaccurate or inconclusive (in 25-35% of cases according to one study [Weiss et al. 2013]).

The second objective would be to build image-based biomarkers to predict the probability of complications. This, in turn, could inform whether surgical intervention is required. Currently, operation decisions are usually informed by heuristic-based scoring methods such as the Spinal Instability Neoplastic Score (SINS) [Fisher et al. 2010] (shown in 8.1) or Tokuhashi score [Tokuhashi et al. 2014]. However, these methods are inherently subjective and a deep-learning model could potentially be more accurate and reproducible.

To pursue these goals, we have acquired an anonymised dataset of 3442 patients referred for whole spine MRIs from multiple NHS trusts across the UK. As well as the images themselves, we have also acquired the associated radiological reports which will allow us to pursue some of the research directions outlined in

Section 8.2.1.

8.2.3 Quantitative Radiological Grading:

A final possible research direction is to develop more objective grading methods for spinal conditions. The current grading methods in SpineNet are quite qualitative. For example, Pfirrmann gradings 3 to 5 are split by the color of disc in a T2 MRI, with the categories being 'grey', 'grey to black' and 'black' respectively. While this can make it easy to quickly grade discs for a human reader, it is clearly subjective, and the large inter-rater variability observed could potentially be explained by different decision boundaries between readers. A possible solution to this would be to also measure more concrete objective features such as disc height and volume, length of herniated disc protrusion, etc. This could be done by first segmenting anatomical structures of interest (e.g. discs, vertebrae, spinal cord, CSF) and then performing quantitative measurements on the resulting segmentation maps. This could provide a more objective basis for assessing spinal degeneration and accordingly may correlate better with patient outcomes.

8.3 Conclusion

In this thesis we explore automated methods of performing radiological grading of clinical spinal MRIs. We also explore the critical role of self-supervision in medical imaging and propose ideas to improve annotation-efficiency of learning methods. It is my sincere hope that this work will contribute, even a little, to a future in which high-quality healthcare monitoring is accessible to everyone, regardless of background or income.

References

Max Aebi (2005). "The adult scoliosis". In: *European Spine Journal*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan (2022). "Flamingo: a Visual Language Model for Few-Shot Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Mohammed Alloulah and Maximilian Arnold (2023). "Look, Radiate, and Learn: Self-Supervised Localisation via Radio-Visual Correspondence". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran (2020). "Self-Supervised Learning by Cross-Modal Audio-Video Clustering". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Cheolhong An, Yiqian Wang, Junkang Zhang, and Truong Q. Nguyen (2022). "Self-Supervised Rigid Registration for Multimodal Retinal Images". In: *IEEE Transactions on Image Processing*.

Relja Arandjelović and Andrew Zisserman (2017). "Look, Listen and Learn". In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Relja Arandjelović and Andrew Zisserman (2018). "Objects That Sound". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi (2020). "Self-labelling via Simultaneous Clustering and Representation Learning". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Reza Azad, Lucas Rouhier, and Julien Cohen-Adad (2021). "Stacked hourglass network with a multi-level attention mechanism: Where to look for intervertebral disc labeling". In: *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021*.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman (2021). "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval". In: *IEEE International Conference on Computer Vision*.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay (2022). "Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Magnus Borga (2018). "MRI adipose tissue and muscle composition analysis—a review of automation techniques". In: *The British Journal of Radiology*.

W. Brinjikji, P.H. Luetmer, B. Comstock, B.W. Bresnahan, L.E. Chen, R.A. Deyo, S. Halabi, J.A. Turner, A.L. Avins, K. James, J.T. Wald, D.F. Kallmes, and J.G. Jarvik (2015). "Systematic Literature Review of Imaging Features of Spinal Degeneration in Asymptomatic Populations". In: *AJNR. American journal of neuroradiology*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language Models are Few-Shot Learners". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Joseph E. Burns, Jianhua Yao, Tatjana S. Wiese, Hector E. Muñoz, Elizabeth C. Jones, and Ronald M. Summers (2013). "Automated Detection of Sclerotic Metastases in the Thoracolumbar Spine at CT". In: *Radiology*.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá (2020). "PadChest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical Image Analysis*.

Yunliang Cai, Said Osman, Manas Sharma, Mark Landis, and Shuo Li (2015). "Multi-Modality Vertebra Recognition in Arbitrary Views Using 3D Deformable Hierarchical Model". In: *IEEE Transactions on Medical Imaging*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (2020). "End-to-End Object Detection with Transformers". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin (2020). "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin (2021). "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Victor Suárez-Paniagua, Richard Tobin, William Whiteley, Honghan Wu, and Beatrice Alex (2021). "A systematic review of natural language processing applied to radiology reports". In: *BMC Medical Informatics and Decision Making*.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman (2014). "Return of the Devil in the Details: Delving Deep into Convolutional Nets". In: *Proceedings of the British Machine Vision Conference (BMVC)*.

Shouyu Chen, Yin Wang, Ke Sun, and Xiwen Sun (2022). "Re-weighting and Hierarchical Pre-training Boost 3D Medical Self-Supervised Representation". In: *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yao Chen, Yuanhan Mo, Aimee Readie, Gregory Ligozio, Indrajeet Mandal, Faiz Jabbar, Thibaud Coroller, and Bartlomiej W Papiez (2023). "VertXNet: An Ensemble Method for Vertebrae Segmentation and Identification of Spinal X-Ray". In: *arXiv preprint arXiv:2302.03476*.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger (2016). "3D U-Net: Learning Dense Volumetric Segmentation

from Sparse Annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Joshua M. Eisenberg and Stuart L. Weinstein (2022). *Adolescent idiopathic scoliosis*. BMJ Best Practice, https://bestpractice.bmj.com/topics/en-gb/979 (accessed 05/04/2023).

P. Felzenszwald, D. McAllester, and D. Ramanan (2008). "A Discriminatively Trained, Multiscale, Deformable Part Model". In: *Proc. Conference on Computer Vision and Pattern Recognition*.

Charles G. Fisher, Christian P. DiPaola, Timothy C. Ryken, Mark H. Bilsky, Christopher I. Shaffrey, Sigurd H. Berven, James S. Harrop, Michael G. Fehlings, Stefano Boriani, Dean Chou, Meic H. Schmidt, David W. Polly, Roberto Biagini, Shane Burch, Mark B. Dekutoski, Aruna Ganju, Peter C. Gerszten, Ziya L. Gokaslan, Michael W. Groff, Norbert J. Liebsch, Ehud Mendel, Scott H. Okuno, Shreyaskumar Patel, Laurence D. Rhines, Peter S. Rose, Daniel M. Sciubba, Narayan Sundaresan, Katsuro Tomita, Peter P. Varga, Luiz R. Vialle, Frank D. Vrionis, Yoshiya Yamada, and Daryl R. Fourney (Oct. 2010). "A Novel Classification System for Spinal Instability in Neoplastic Disease: An Evidence-Based Approach and Expert Consensus From the Spine Oncology Study Group". In: *Spine*.

Daniel Forsberg, Erik Sjöblom, and Jeffrey L. Sunshine (2017). "Detection and Labeling of Vertebrae in MR Images Using Deep Learning with Clinical Annotations as Training Data". In: *Journal of Digital Imaging*.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid (2020). "Multi-modal Transformer for Video Retrieval". In: *European Conference on Computer Vision (ECCV)*.

Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu (2019). "DermGAN: Synthetic Generation of Clinical Skin Images with Pathology". In: *Machine Learning for Health NeurIPS Workshop*.

Somoballi Ghoshal, Sourav Banu, Amlan Chakrabarti, Susmita Sur-Kolay, and Alok Pandit (2019). "3D reconstruction of spine image from 2D MRI slices along one axis". In: *IET Image Processing*.

Ben Glocker, J. Feulner, Antonio Criminisi, David R. Haynor, and E. Konukoglu (2012). "Automatic Localization and Identification of Vertebrae in Arbitrary Field-of-View CT Scans". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Ben Glocker, Darko Zikic, Ender Konukoglu, David R. Haynor, and Antonio Criminisi (2013). "Vertebrae Localization in Pathological Spine CT via Dense Classification from Sparse Annotations". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the International Conference on Machine Learning (ICML)*.

Charles Greenough (2016). *The National Back Pain Pathway*. https://www.england.nhs.uk/blog/charles-greenough/. Accessed on April 17, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko (2020). "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger (2017). "On Calibration of Modern Neural Networks". In: *International Conference on Machine Learning*.

A. Gupta, A. Vedaldi, and A. Zisserman (2018). "Inductive Visual Localisation: Factorised Training for Superior Generalisation". In: *Proc. British Machine Vision Conference*.

Michael U. Gutmann and Aapo Hyvärinen (2012). "Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics". In: *Journal of Machine Learning Research*.

Matthias Hammon, Peter Dankerl, Alexey Tsymbal, Michael Wels, Michael Kelm, Matthias May, Michael Suehling, Michael Uder, and Alexander Cavallaro (2013). "Automatic detection of lytic and blastic thoracolumbar spine metastases on computed tomography". In: *European Radiology*.

Tengda Han, Weidi Xie, and Andrew Zisserman (2020). "Self-supervised Co-training for Video Representation Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

J. Hays, M. Leordeanu, A. Efros, and Y. Liu (2006). "Discovering Texture Regularity as a Higher-Order Correspondence Problem". In: *Proc. European Conference on Computer Vision*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (2021). "Masked Autoencoders Are Scalable Vision Learners". In: *ArXiv preprint, arXiv:2111.06377*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick (2020). "Momentum Contrast for Unsupervised Visual Representation Learning". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Michael Brady, and Julia A. Schnabel (2012). "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Olivier Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, Ali Eslami, and Aaron van den Oord (2020). "Data-Efficient Image Recognition with Contrastive Predictive Coding". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Richard Herzog, Daniel R. Elgort, Adam E. Flanders, and Peter J. Moley (2017). "Variability in diagnostic error rates of 10 MRI centers performing lumbar spine MRI examinations on the same patient within a 3-week period". In: *Case Reports Spine J.*

Jeremy Howard and Sebastian Ruder (2018). "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung (2021). "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition". In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt (2021). *OpenCLIP*.

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

A. Jain, J. Tompson, M. Andriluka, G. Taylor, and C. Bregler (2014). "Learning Human Pose Estimation Features with Convolutional Networks". In: *Proc. International Conference on Learning Representations*.

Amir Jamaludin, Jeremy Fairbank, Ian Harding, Timor Kadir, Tim Peters, Andrew Zisserman, and Emma Clark (2020a). "Identifying Scoliosis in Population-Based Cohorts: Automation of a Validated Method Based on Total Body Dual Energy X-ray Absorptiometry Scans". In: *Calcified Tissue International*.

Amir Jamaludin, Timor Kadir, Emma Clark, and Andrew Zisserman (2018). "Predicting Scoliosis in DXA Scans Using Intermediate Representations". In: *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*.

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

Amir Jamaludin, Timor Kadir, and Andrew Zisserman (2015). "Automatic Modic Changes Classification in Spinal MRI New words". In: *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*.

Amir Jamaludin, Timor Kadir, and Andrew Zisserman (2017a). "Self-Supervised Learning for Spinal MRIs". In: *MICCAI Workshop on Deep Learning in Medical Image Analysis*.

Amir Jamaludin, Timor Kadir, and Andrew Zisserman (2017b). "SpineNet: Automated Classification and Evidence Visualization in Spinal MRIs". In: *Medical Image Analysis*.

Amir Jamaludin, Timor Kadir, Andrew Zisserman, Ian McCall, Frances Williams, Gail Lang, Elaine Buchanan, Jill Urban, and Jeremy Fairbank (Mar. 2023). "ISSLS

PRIZE in Clinical Science 2023: comparison of degenerative MRI features of the intervertebral disc between those with and without chronic low back pain. An exploratory study of two large female populations using automated annotation". In: *European Spine Journal*.

Amir Jamaludin, Meelis Lootus, Timor Kadir, Andrew Zisserman, Jill Urban, Michele C. Battié, Jeremy Fairbank, and Ian McCall (2017c). "Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist". en. In: *European Spine Journal*.

Amir Jamaludin, Meelis Lootus, Timor Kadir, Andrew Zisserman, Jill Urban, Michele Battié, Jeremy Fairbank, and Iain McCall (2017d). "Automation of Reading of Radiological Features from Magnetic Resonance Images (MRIs) of the Lumbar Spine Without Human Intervention Is Comparable with an Expert Radiologist". In: *European Spine Journal*.

Amir Jamaludin, Rhydian Windsor, Sarim Ather, Timor Kadir, Andrew Zisserman, Jurgen Braun, Lianne Gensler, Pedro Machado, Mikkel Ostergaard, Denis Poddubnyy, Thibault Coroller, Brian Porter, Shepard Mpofu, and Aimee Readie (2020b). "Machine learning based Berlin scoring of magnetic resonance images of the spine in patients with ankylosing spondylitis from the Measure 1 study." In: *Annals of the Rheumatic Diseases*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung (2023). "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig (2021). "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision". In: *Proceedings of the International Conference on Machine Learning (ICML)*.

Alistair E. W. Johnson et al. (Dec. 2019). "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data*.

Maxime Kayser, Cornelius Emde, Oana Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz (2022). "Explaining Chest X-ray Pathologies in Natural Language". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Sarah E. Kerr, Catherine A. Schnabel, Peggy S. Sullivan, Yi Zhang, Veena Singh, Brittany Carey, Mark G. Erlander, W. Edward Highsmith, Sarah M. Dry, and Elena F. Brachtel (2012). "Multi-site Validation Study to Determine Performance

Characteristics of a 92-Gene Molecular Cancer ClassifierMultisite Validation of a 92-Gene Molecular Cancer Classifier". In: *Clinical Cancer Research*.

Diederik Kingma and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Per Kjaer, Charlotte Leboeuf-Yde, Lars Korsholm, Joan Solgaard Sorensen, and Tom Bendix (2005). "Magnetic resonance imaging and low back pain in adults: a diagnostic imaging study of 40-year-old men and women". In: *Spine*.

Bruno Korbar, Du Tran, and Lorenzo Torresani (2018). "Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Niina Korpinen, Petteri Oura, Tiina Väre, Markku Niskanen, Jaakko Niinimäki, Jaro Karppinen, and Juho-Antti Junno (2020). "Temporal Trends in Vertebral Dimensions – a case study from Finland". en. In: *Scientific Reports*.

Narasimharao Kowlagi, Huy Hoang Nguyen, Terence McSweeney, Simo Saarakkala, Jaro Karppinen, and Aleksei Tiulpin (2022). "A Stronger Baseline For Automatic Pfirrmann Grading Of Lumbar Spine MRI Using Deep Learning". In: *arXiv preprint arXiv:2210.14597*.

Thomas Leung and Jitendra Malik (1996). "Detecting, localizing and grouping repeated scene elements from an image". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kai-Uwe Lewandrowskl, Narendran Muraleedharan, Steven Allen Eddy, Vikram Sobti, Brian D. Reece, Jorge Felipe ramírez León, and Sandeep Shah (2022). "Feasibility of Deep Learning Algorithms for Reporting in Routine Spine Magnetic Resonance Imaging". In: *International Journal of Spine Surgery* (s3).

Jun Li, Quan Quan, and S. Kevin Zhou (2022). "MixCL: Pixel label matters to contrastive learning". In: *ArXiv preprint, arXiv:2203.02114*.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan (2022). "Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M. Wells (2021). "Multimodal Representation Learning via Maximization of Local Mutual Information". In:

Meelis Lootus, Timor Kadir, and Andrew Zisserman (2013). "Vertebrae Detection and Labelling in Lumbar MR Images". In: *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*.

Meelis Lootus, Timor Kadir, and Andrew Zisserman (2014a). "Radiological Grading of Spinal MRI". In: *MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging*.

Meelis Lootus, Timor Kadir, and Andrew Zisserman (2014b). "Vertebrae Detection and Labelling in Lumbar MR Images". en. In: *Computational Methods and Clinical Applications for Spine Imaging*.

David Lowe (Sept. 1999). "Object recognition from local scale-invariant features". In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

David Lowe (2004). "Distinctive Image Features from Scale-Invariant Keypoints". In: *IJCV*.

Jen-Tang Lu, Stefano Pedemonte, Bernardo Bizzo, Sean Doyle, Katerine P. Andriole, Mark H. Michalski, Gilberto Gonzalez, and Stuart R. Pomerantz (2018). "Deep Spine: Automated Lumbar Vertebral Segmentation, Disc-Level Designation, and Spinal Stenosis Grading using Deep Learning". In: *Proceedings of the Machine Learning For Healthcare Conference (MLHC)*.

Giulio Maccauro, Maria Silvia Spinelli, Sigismondo Mauro, Carlo Perisano, Calogero Graci, and Michele Attilio Rosa (2011). "Physiopathology of Spine Metastasis". In: *International Journal of Surgical Oncology*.

David Mattes, David R. Haynor, Hubert Vesselle, Thomas K. Lewellyn, and William Eubank (2001). "Nonrigid multimodality image registration". In: *Medical Imaging 2001: Image Processing*. Ed. by Milan Sonka and Kenneth M. Hanson. International Society for Optics and Photonics.

Terence McSweeney, Aleksei Tiulpin, Simo Saarakkala, Jaakko Niinimäki, Rhydian Windsor, Amir Jamaludin, Timor Kadir, Jaro Karppinen, and Juhani Määttä (2022). "External Validation of SpineNet, an Open-source Deep Learning Model for Grading Lumbar Disc Degeneration MRI Features, Using the Northern Finland Birth Cohort 1966". In: *Spine*.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer (2022). "TrackFormer: Multi-Object Tracking with Transformers". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zamir Merali, Justin Z. Wang, Jetan H. Badhiwala, Christopher D. Witiw, Jefferson R. Wilson, and Michael G. Fehlings (2021). "A deep learning model for

detection of cervical spinal cord compression in MRI scans". en. In: *Scientific Reports*.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman (2021). "Thinking Fast and Slow: Efficient Text-to-Visual Retrieval with Transformers". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie (2021). "SLIP: Self-supervision meets Language-Image Pre-training". In: *arXiv preprint arXiv:2112.12750*.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert (2022). "Joint Learning of Localized Representations from Medical Images and Reports". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

National Institute for Health and Care Excellence (NICE) (2022). *Back pain - low (without radiculopathy)*. NICE Clinical Knowledge Summaries, https://cks.nice.org.uk/topics/back-pain-low-without-radiculopathy/. Accessed on April 17, 2023.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer (2021). "Progressive Transformer-Based Generation of Radiology Reports". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding". In: *arxiv preprint, arXiv:1807.03748*.

OpenAI (2023). *GPT-4*.

J.A. Ortiz Gomez (1995). "The incidence of vertebral body metastases". In: *International Orthopaedics*.

Andrew Owens and Alexei A. Efros (2018). "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Cagatay Ozturk, Selhan Karadereler, Ibrahim Ornek, Meric Enercan, Kursat Ganiyusufoglu, and Azmi Hamzaoglu (2010). "The role of routine magnetic resonance imaging in the preoperative evaluation of adolescent idiopathic scoliosis". In: *International Orthopaedics*.

Saba Pasha and John Flynn (2018). “Data-driven Classification of the 3D Spinal Curve in Adolescent Idiopathic Scoliosis with an Applications in Surgical Outcome Prediction”. en. In: *Scientific Reports*.

C. Payer, D. Štern, H. Bischof, and M. Urschler (2016). “Regressing Heatmaps for Multiple Landmark Localization Using CNNs”. en. In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Z. Peng, J. Zhong, W. Wee, and J. Lee (2005). “Automated Vertebra Detection and Segmentation from the Whole Spine MR Images”. In: *IEEE Engineering in Medicine and Biology 27th Annual Conference*.

Christian W. A. Pfirrmann, Alexander Metzdorf, Marco Zanetti, Juerg Hodler, and Norbert Boos (2001). “Magnetic Resonance Classification of Lumbar Intervertebral Disc Degeneration:” in: *Spine*.

T. Pfister, J. Charles, and A. Zisserman (2015). “Flowing ConvNets for Human Pose Estimation in Videos”. In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui (2021). “Spatiotemporal Contrastive Video Representation Learning”. In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the International Conference on Machine Learning (ICML)*.

Alec Radford, Jong Wook Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI.

Bernardo Romera-Paredes and Philip Torr (2016). “Recurrent Instance Segmentation”. In: *Proc. European Conference on Computer Vision*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

L. Scheibenreif, M. Mommert, and D. Borth (2022). “Contrastive Self-Supervised Data Fusion For Satellite Imagery”. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Harald Scheidl, Stefan Fiel, and Robert Sablatnig (2018). “Word Beam Search: A Connectionist Temporal Classification Decoding Algorithm”. In: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*.

Elad Segal, Ben Bogin, and Jonathan Berant (2022). “Training Vision-Language Models with Less Bimodal Supervision”. In: *4th Conference on Automated Knowledge Base Construction*.

Brian Shaw, Frederick L. Mansfield, and Lawrence Borges (1989). “One-stage posterolateral decompression and stabilization for primary and metastatic vertebral tumors in the thoracic and lumbar spine”. In: *Journal of Neurosurgery*.

Karen Simonyan, Andrew Zisserman, and Antonio Criminisi (2011). “Immediate Structured Visual Search for Medical Images”. In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins (2015a). “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*.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins (Mar. 2015b). “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*.

Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi (2021). “Multimodal Self-supervised Learning for Medical Image Analysis”. In: *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert (2020). “3D Self-Supervised Methods for Medical Imaging”. In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Rong Tao and Guoyan Zheng (2021). “Spine-Transformers: Vertebra Detection and Localization in Arbitrary Field-of-View Spine CT with Transformers”. In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Austin Tapp, Michel Audette, and James Bennett (2021). “Automatic Generation and Novel Validation of Patient-Specific, Anatomically Inclusive Scoliosis Models for Biomechanics-Informed Surgical Planning”. In: *College of Engineering & Technology (Batten) Posters*.

Hilary J. Taylor, Ian Harding, John Hutchinson, Ian Nelson, Ashley Blom, Jon H. Tobias, and Emma M. Clark (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*.

Bernhard Tins and Birender Balain (2016). “Incidence of numerical variants and transitional lumbosacral vertebrae on whole-spine MRI”. In: *Insights into Imaging*.

Matthew Toews, Lilla Zöllei, and William M. Wells (2013). “Feature-Based Alignment of Volumetric Multi-modal Images”. In: *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*.

Yasuaki Tokuhashi, Hiroshi Uei, Masashi Oshima, and Yasumitsu Ajiro (2014). “Scoring system for prediction of metastatic spine tumor prognosis”. In: *World journal of orthopedics*.

Johnatan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler (2014). “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”. In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Floris R. van Tol, Julie R. A. Massier, Geert W. J. Frederix, F. Cumhur Öner, Helena M. Verkooijen, and Jorrit-Jan Verlaan (2021a). “Costs Associated With Timely and Delayed Surgical Treatment of Spinal Metastases”. en. In: *Global Spine Journal*. (Visited on 06/17/2021).

Floris R. van Tol, Anne L. Versteeg, Helena M. Verkooijen, F. Cumhur Öner, and Jorrit-J Verlaan (2021b). “Time to Surgical Treatment for Metastatic Spinal Disease: Identification of Delay Intervals”. en. In: *Global Spine Journal*. (Visited on 06/17/2021).

Paul Viola and William Wells (1995). “Alignment by Maximization of Mutual Information”. In: *Proceedings of the International Conference on Computer Vision (ICCV)*.

Fakai Wang, Kang Zheng, Le Lu, Jing Xiao, Min Wu, and Shun Miao (2021). "Automatic vertebra localization and identification in CT by spine rectification and anatomically-constrained optimization". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Juan Wang, Zhiyuan Fang, Ning Lang, Huishu Yuan, Min-Ying Su, and Pierre Baldi (2017). "A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks". In: *Computers in biology and medicine*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers (2018). "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays". In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jason Wei and Kai Zou (2019). "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lawrence M. Weiss, Peiguo Chu, Brock E. Schroeder, Veena Singh, Yi Zhang, Mark G. Erlander, and Catherine A. Schnabel (2013). "Blinded comparator study of immunohistochemical analysis versus a 92-gene cancer classifier in the diagnosis of the primary site in metastatic tumors". In: *The Journal of Molecular Diagnostics*.

Rhydian Windsor and Amir Jamaludin (2020). "The Ladder Algorithm: Finding Repetitive Structures in Medical Images by Induction". In: *IEEE International Symposium on Biomedical Imaging*.

Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman (2020). "A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman (2021). "Self-Supervised Multi-Modal Alignment for Whole Body Medical Imaging". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman (2022). "SpineNetV2: Automated Detection, Labelling and Radiological Grading Of Clinical MR Scans". In: *Technical Report, arXiv:2205.01683*.

World Health Organization (2022). *WHO Factsheet on Musculoskeletal Health*. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions. Accessed: April 17, 2023.

Dong Yang, Tao Xiong, Daguang Xu, Qiangui Huang, David Liu, S. Kevin Zhou, Zhoubing Xu, Jin-Hyeong Park, Mingqing Chen, Trac D. Tran, Sang Peter Chin, Dimitris Metaxas, and Dorin Comaniciu (2017). "Automatic Vertebra Labeling in Large-Scale 3D CT Using Deep Image-to-Image Network with Message Passing and Sparsity Regularization". In: *Information Processing in Medical Imaging*.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao (2022). "Knowledge matters: Chest radiology report generation with general and specific knowledge". In: *Medical Image Analysis*.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu (2021). "AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation". In: *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Alexey Zakharov, Maxim Pisov, Alim Bukharaev, Alexey Petraikin, Sergey Morozov, Victor Gombolevskiy, and Mikhail Belyaev (2023). "Interpretable vertebral fracture quantification via anchor-free landmarks localization". In: *Medical Image Analysis*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz (2022). "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *Proc. MLHC*.

Shen Zhao, Bin Chen, Heyou Chang, Bo Chen, and Shuo Li (2022). "Reasoning discriminative dictionary-embedded network for fully automatic vertebrae tumor diagnosis". In: *Medical Image Analysis*.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl (2019). "Objects as Points". In: *arXiv preprint arXiv:1904.07850*.

Dženan Zukić, Aleš Vlasák, Jan Egger, Daniel Hořínek, Christopher Nimsky, and Andreas Kolb (2014). "Robust Detection and Segmentation for Diagnosis of Vertebral Diseases Using Routine MR Images". In: *Computer Graphics Forum*.

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 "The Ladder Algorithm: Finding Repetitive Structures in Medical Images by Induction" in Chapter 2.

Paper title	The Ladder Algorithm: Finding Repetitive Structures in Medical Images by Induction
Authors	**Rhydian Windsor**, Amir Jamaludin
Publication status	Published
Publication details	International Symposium on Biomedical Imaging (ISBI), 2020.

Student Confirmation

Student name	Rhydian Windsor	
Contribution to the paper	First-author contribution: • conception of research ideas (joint with other author) • design and implementation of models • running experiments on all datasets • writing and presentation of the paper	
Signature and Date		

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		

Statement of Authorship for the paper "A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI" in Chapter 3.

Paper title	A Convolutional Approach to Vertebrae Detection and Labelling in Whole Spine MRI
Authors	**Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman
Publication status	Published
Publication details	Medical Image Computing and Computer Aided Intervention (MICCAI), 2020. Early Acceptance

Student Confirmation

Student name	Rhydian Windsor	
Contribution to the paper	First-author contribution: • conception of research ideas (joint with other authors) • design and implementation of models • writing and presentation of the paper	
Signature and Date		

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		

Statement of Authorship for the paper "Self-Supervised Multi-Modal Alignment for Whole Body Medical Imaging" in Chapter 5.

Paper title	Self-Supervised Multi-Modal Alignment for Whole Body Medical Imaging
Authors	**Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman
Publication status	Published
Publication details	Medical Image Computing and Computer Aided Intervention (MICCAI), 2021.

Student Confirmation

Student name	Rhydian Windsor	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models and experiments • writing and presentation of the paper	
Signature and Date		

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		

Statement of Authorship for the paper "3D Spinal Column Segmentation with Single Plane 2D-Projected Annotations" in Chapter 6.

Paper title	3D Spinal Column Segmentation with Single Plane 2D-Projected Annotations
Authors	**Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman
Publication status	Peer Reviewed, Not Published
Publication details	Accepted as an extended abstract (non-archival) at Medical Imaging Meets Neurips (MedNeurips), 2021.

Student Confirmation

Student name	Rhydian Windsor	
Contribution to the paper	First-author contribution: • initial conception of research idea • design and implementation of models and experiments • writing and presentation of the paper	
Signature and Date		

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		

Statement of Authorship for the paper "Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading" in Chapter 4.

Paper title	Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading
Authors	**Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman
Publication status	Published
Publication details	Medical Image Computing and Computer Aided Intervention (MICCAI), 2022. Early Acceptance.

Student Confirmation

Student name	Rhydian Windsor	
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		

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		

Statement of Authorship for the paper ""Vision-Language Modelling for Radiological Imaging and Reports in the Low Data Regime" in Chapter 7.

Paper title	"Vision-Language Modelling for Radiological Imaging and Reports in the Low Data Regime
Authors	**Rhydian Windsor**, Amir Jamaludin, Timor Kadir, Andrew Zisserman
Publication status	Published
Publication details	Medical Imaging with Deep Learning (MIDL) 2023 (Oral presentation).

Student Confirmation

Student name	Rhydian Windsor	
Contribution to the paper	First-author contribution: • conception of research ideas • design and implementation of models and experiments • writing and presentation of the paper	
Signature and Date		

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		