VertFault: 3D Segmentation from CT Images for Vertebral Fracture Detection

Prakash Nair

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2024

Abstract

Vertebral fractures are a significant medical concern due to their impact on mobility, chronic pain, and overall quality of life, particularly in older populations where osteoporosis is prevalent. Despite the critical nature of early detection, a large proportion of vertebral fractures go undiagnosed due to asymptomatic presentations and the manual nature of current diagnostic methods, which are time-consuming and prone to human error. This project, "VertFault: 3D Segmentation from CT Images for Vertebral Fracture Detection," addresses these challenges by developing an automated tool for detecting and grading vertebral fractures using CT scans. The research employs a two-stage pipeline: a U-Net-based model for vertebra localisation and a multi-task learning framework for 3D segmentation and fracture classification. While the model shows robustness in vertebra localisation, fracture detection in a class-imbalanced setting remains challenging, particularly in accurately identifying higher-grade fractures. This study highlights the potential of automated tools in improving clinical outcomes, though further research is necessary to enhance fracture detection accuracy.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Prakash Nair*)

Acknowledgements

I would like to express my deepest gratitude to **Dr. Changjian Li**, **Sonia Dahdouh**, and **Keith Goatman**. Their unwavering support, guidance, and encouragement have been instrumental throughout this process. Thank you for believing in me and for your profound contributions to this work.

Table of Contents

Chapter 1

Introduction

The evolution of spinal imaging, driven by advancements in computed tomography (CT), has significantly improved the diagnosis and treatment of spinal pathologies, from degenerative diseases to traumatic injuries [1]. CT scans play a crucial role in spinal diagnosis due to their ability to provide high-resolution, detailed cross-sectional images of the spine. These scans offer superior bone contrast, making it possible to detect subtle fractures and other bony abnormalities with high precision. The detailed imagery obtained from CT scans enables clinicians to assess the extent of vertebral fractures, determine the involvement of surrounding structures, and plan appropriate interventions.

However, the accurate diagnosis of vertebral fractures remains a major challenge, with two-thirds of such fractures going undetected [2]. This under-diagnosis is often due to asymptomatic presentations or non-specific symptoms, which can lead to severe consequences as patients with undiagnosed fractures are at a higher risk of subsequent fractures, resulting in chronic pain and diminished quality of life [3].

The complexity of spinal anatomy and subtle fracture manifestations on imaging further complicate diagnosis. Traditional methods rely on radiologists' expertise to meticulously analyse scans, a process that is both time-consuming and susceptible to human error [4].

Recent advancements in computer vision (CV), such as Convolutional Neural Networks and consequently, U-Net architectures, have shown significant promise in automating and enhancing medical image analysis, achieving remarkable accuracy in tasks such as tumor detection [5], organ segmentation [6], and disease classification [7]. Applying such methods to the field of spinal imaging promises similarly positive advancements in the detection and diagnosis of vertebral fractures, making the process

more efficient, accurate, and less reliant on individual clinical expertise and manual analysis.

This project focuses on refining the tasks of automated localisation and segmentation of vertebrae to develop a robust tool for the end goal of detection and grading of vertebral fractures in 3D CT scans. Positioned at the intersection of medical imaging, artificial intelligence, and clinical orthopaedics, this endeavor holds promise in enhancing clinical outcomes by providing an effective tool for the early detection and grading of vertebral fractures, facilitating timely and targeted interventions.

Chapter 2

Background

2.1 Spinal Anatomy and Pathologies

2.1.1 Anatomy of the Spine

To lay the preliminary understanding for our project, we delve into the detailed anatomy of the spine in order to contextualise the terminology employed in later sections. The spine consists of 33 stacked vertebra that form the spinal column allowing for body movement whilst protecting the spinal cord [8]. The S-shape of the spine consists of three main curvatures: the cervical (neck) and lumbar (low back) regions have a concave curvature, while the thoracic region has a convex curve. This curved shape provides structural support and flexibility, allowing for upright posture while effectively distributing weight and absorbing shock during movement. This curvature also protects the spinal cord by reducing the risk of direct impact from external forces.

There are 5 distinct regions of the spine: cervical, thoracic, lumbar, sacral and coccyx [8]. The lowest regions of the sacrum and coccyx are immobile, and as such, only the upper 24 vertebrae are movable. Figure 2.1 visualises the regions in greater detail including their alphanumerical naming [9].

The vertebrae themselves are individual bony units that together form the vertebral column, providing support and protection for the spinal cord. Each vertebra consists of three main components: the vertebral body, vertebral arch, and processes. The vertebral body, positioned anteriorly, provides the primary weight-bearing structure. The vertebral arch, located posteriorly, encloses the spinal cord and forms the vertebral foramen, the opening that collectively constitute the spinal canal. Extending from the arch are processes: one that points posteriorly and two on either side which jointly serve

as attachment sites for muscles and ligaments [10].

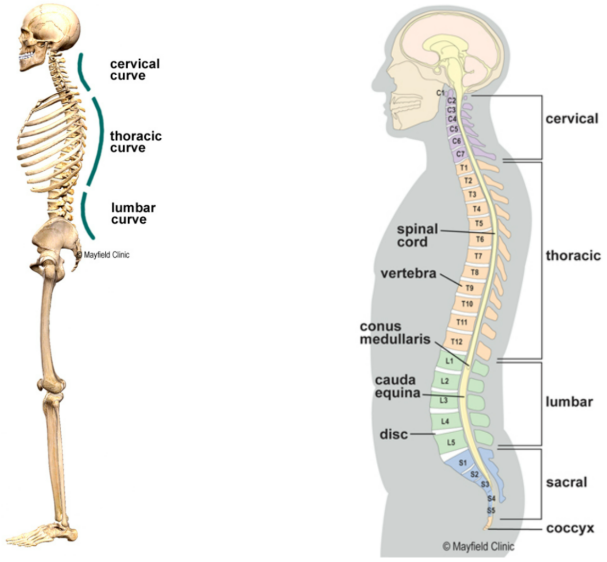

Figure 2.1: (Left) Typical S-shape curvature of the spine. (Right) 5 regions of the spine with anatomical labels. Adapted from [9].

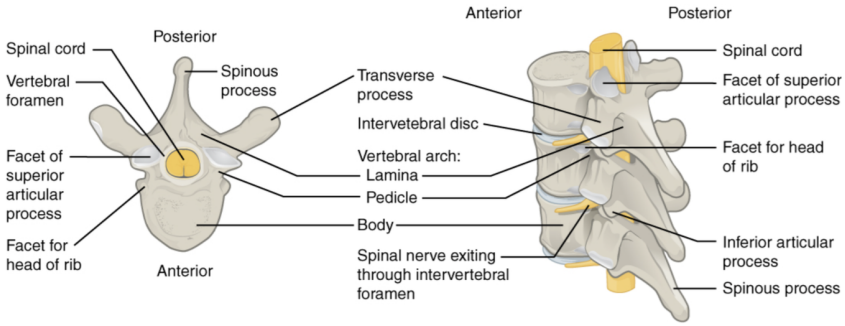

Figure 2.2: Vertebral structures present throughout the spinal column. Adapted from [11].

2.1.2 Vertebral Fractures: Causes, Symptoms, and Implications

Vertebral fractures are defined as a break in one of the vertebrae in the spine [12] and are significant due to their impact on mobility and quality of life. Compression fractures, common in osteoporosis and osteopenia, involve the collapse of the vertebral body, often due to weakened bone. Burst fractures result from severe trauma, causing the vertebra to shatter. Other types include flexion-distraction fractures from sudden

forward flexion, and fracture-dislocations where bone fragments separate and misalign the spine.

Osteoporosis is the condition characterised by diminished bone density and structural deterioration, rendering bones fragile and prone to fractures. As individuals age, the risk of osteoporosis increases, with postmenopausal women being particularly susceptible due to the decline in estrogen levels essential for maintaining bone density [13]. Osteopenia, a precursor to osteoporosis, involves lower than normal reference bone density but not to the extent of osteoporosis. Both conditions significantly elevate the risk of fractures, particularly vertebral fractures, as the reduced bone density weakens the structure of vertebrae [14]. These fractures can lead to chronic pain, reduced mobility, and further complications, highlighting the importance of early detection and intervention in patients with osteopenia and osteoporosis to prevent progression and mitigate the risk of severe fractures.

In the older population, osteoporotic compression fractures are the most common type of vertebral fracture [15]. Clinical manifestations of such fractures can often be subtle, leading to a high rate of undiagnosed cases [2]. Many vertebral fractures are asymptomatic or occur in the absence of specific trauma. When symptoms do appear, they typically include a sudden onset of back pain, loss of height, and spinal deformities like kyphosis [16]. The implications of undiagnosed vertebral fractures are profound, leading to chronic pain, reduced mobility, and a decreased quality of life. More critically, undiagnosed fractures significantly increase the risk of subsequent fractures, including hip fractures, which are associated with high morbidity and mortality [13]. Given such consequences, the importance of early detection and treatment of vertebral fractures cannot be overstated.

Notably, vertebral fractures also demonstrate a bimodal distribution, with younger patients sustaining such fractures due to high-energy traumatic mechanisms (falls from height, vehicular accidents, etc.) [17]. However, such fractures are associated with improved rates of diagnosis due to the traumatic nature of the injury [18]. Despite the higher diagnosis rates in younger individuals due to apparent trauma, there are still instances of missed diagnoses. Improved diagnostic measures are essential for both the young and the elderly to prevent the aforementioned consequences of undiagnosed vertebral fractures.

2.2 Computed Tomography (CT) Imaging for Spinal Pathology

Computed Tomography (CT) is one of the most advanced and widely used imaging modalities in medical diagnostics. CT scanning involves the use of X-rays combined with computer technology to produce detailed cross-sectional images of the body. The X-rays rotate around the patient, and the data collected is computationally reconstructed into slices, providing a detailed view of the internal structure [19]. This method is particularly effective in visualising bone structures and detecting fractures due to its high-resolution 3D images with radiodensity contrast that highlights bony tissues [20].

Within the context of spinal imaging and spinal pathology diagnosis, CT is often considered the superior imaging modality [20]. This preference stems from CT's ability to produce high-resolution images of bone structures in three planes, which is essential for accurately identifying fractures [21]. CT scans are performed more frequently than other imaging modalities [22] due to their speed, cost-effectiveness, and superior ability to visualise bone structures. Pertinently, this higher usage rate results in a large number of publicly available annotated CT datasets, which are essential for developing and training automated diagnostic tools [23]. Furthermore, the contrast between bone and soft tissue provided by CT scans facilitates the development of advanced algorithms for automated detection, enabling precise analysis of the vertebral anatomy.

2.3 Challenges in Vertebral Fracture Diagnosis

Vertebral fractures are primarily diagnosed using CT scans, which provide detailed cross-sectional images of the spine. The diagnosis process typically involves semi-quantitative methods that combine morphometric and manual visual assessment [24]. The preferred semi-quantitative method, proposed by Genant et al. [25], is considered the gold standard due to its accuracy, reliability, and continuous use in clinical studies [26]. The scheme categorises fracture severity based on the extent of anterior, posterior or middle height reduction in vertebrae. The grading system includes three grades: mild (grade 1), with a 20-25% reduction in anterior, middle, or posterior vertebral height; moderate (grade 2), with a 25-40% reduction; and severe (grade 3), with more than a 40% reduction in vertebral height. Figure 2.3 illustrates the grading scheme for the case of anterior height loss, H_a.

Despite the existence of standardised methods of fracture diagnosis and grading,

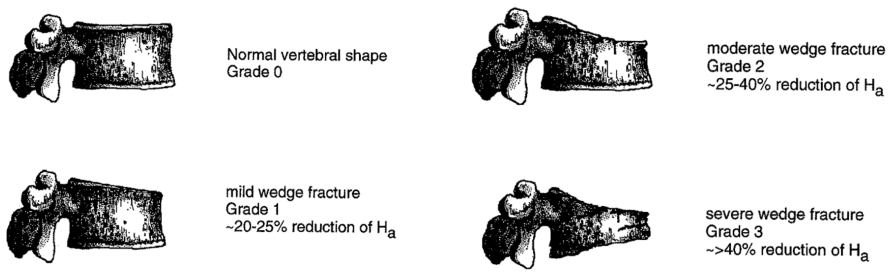

Figure 2.3: Semi-Quantitative grading of vertebral fractures illustrating the the reduction in anterior height. Reduction in the posterior or middle of the vertebra can be assessed using the same scheme. Taken from [25].

there remain existing challenges to diagnosis of vertebral fractures due to various factors. Anatomical variations among patients can impact the accuracy of diagnosis, as variations in vertebral shape and non-fracture related deformities may resemble or obscure true fractures [27]. Subtle manifestations of fractures, such as minor compression or slight endplate deformities, are often difficult to detect visually, even with high-resolution imaging techniques. This subtlety necessitates meticulous manual analysis, which is both time-consuming [4] and prone to human error due to lack of standardisation in the radiologic interpretation [28].

Furthermore, vertebral fractures are commonly present on imaging obtained for other reasons in patients who may not show signs or symptoms suggestive of fracture [28]. As aforementioned, the high rate of under-diagnosis, often due to asymptomatic fractures or those occurring without specific trauma, has severe consequences for patient health, further highlighting the use case and value of automated methods for fracture detection [29].

2.4 ML/AI in Spinal Imaging

Existing research on automated spine analysis corroborates the importance of accurate vertebrae localisation, identification, and segmentation in downstream orthopaedic tasks, such as fracture detection and grading. However, previous work has predominantly focused on the former two tasks. Schmidt et al. [30] employ a classification tree-based approach toward localisation and identification, which incorporates appearance and geometric relationships of spine parts by using local feature vectors from sub-volumes of the image to predict the probabilities of specific image points being an intervertebral

disc. Glocker et al. [31] developed a two-stage method for localising and identifying vertebrae in CT scans. Initially, regression forests predict vertebra centroids using supervised learning with image point features and displacements. Precision is improved by incorporating long-range spatial features, which use context from surrounding organs, enabling accurate center estimation. A Hidden Markov Model (HMM) then refines these predictions, accounting for the sequential and probabilistic relationships of vertebrae, correcting misalignments, and aligning with the spine's overall shape.

Recent advances have shifted towards deep learning, particularly Convolutional Neural Network (CNN) based methods, for enhanced vertebrae localisation and identification in CT scans. Liao et al. [32] introduced a multi-task 3D CNN to extract short-range contextual information from vertebrae samples, paired with a Bidirectional Recurrent Neural Network (Bi-RNN) [33] to analyse long-range spatial relationships along the spine. This approach uses a deep multi-task 3D CNN, converted into a fully convolutional network (FCN) to accommodate various CT image sizes, and a Bi-RNN to refine spatial data from the FCN, improving anatomical accuracy of the spine model. For the same task, Cui et al. [34] present a novel framework that improves vertebrae identification through a module that captures both the upward and downward relationships between vertebrae and a continuous vertebrae label map, as opposed to one that is discrete. This model not only addresses the global structure of the spine but also captures intricate local details of each vertebra. The approach employs a localisation network to create a Gaussian-like 3D heatmap and an offset map, which together refine vertebra center predictions using a unique voting scheme that incorporates Chamfer distance supervision to ensure proximity to actual vertebra positions.

Within the frame of medical image segmentation, the U-Net architecture has proved to be a pivotal advancement. The U-Net is characterised by its symmetric, U-shaped structure, designed to capture both local and contextual information efficiently, which is essential for tasks like vertebrae localisation and segmentation [35] [36]. Payer et al. [37] propose a U-Net based, multi-stage method for vertebrae localisation and segmentation in CT images, utilising a progressive approach to enhance precision. Initially, a coarse localisation of the spine is achieved through a U-Net architecture. This is followed by a detailed localisation and identification of individual vertebrae using a fully-connected CNN that merges local landmark appearances with their spatial configurations to facilitate heatmap regression for accurate localisation. Finally, another U-Net performs binary segmentation of each identified vertebrae in a high resolution, before merging the individual predictions into the resulting multi-label vertebrae segmentation.

With regard to the task of automated vertebral fracture detection in CT scans, relatively few works have addressed the problem comprehensively. However, with the aforementioned advancements in computer vision and increased concern regarding the consequences of undiagnosed fractures, there has been a growing focus on developing more robust and precise methods for fracture detection and grading. Zhang et al. [38] propose a multistage ensemble framework, starting with a U-Net and Graph Convolutional Network (U-GCN) [39] for locating and identifying vertebrae in the thoracic and lumbar sections of the spine, followed by a classification network to detect vertebral fractures in regions-of-interest cropped around the localised vertebrae. Nadeem et al. [40] developed a chest CT-based automated method for fracture assessment. Their approach begins with the computation of a voxel-level vertebral body likelihood map from chest CT scans using a trained deep learning network. To address the challenge of fused vertebrae in CT images, intensity autocorrelation is employed for separation. Vertebral heights are then computed using contour analysis on the central anterior-posterior plane of each vertebral body. Finally, vertebral fracture status is assessed using ratio functions of vertebral heights.

Despite these advancements, existing methods have several limitations. They often treat vertebral localisation, segmentation, and fracture detection/grading as separate tasks, which can lead to inefficiencies and suboptimal performance due to the lack of shared information between tasks. Additionally, these methods may struggle with accurately capturing the detailed morphology of vertebrae and the subtle variations associated with fractures.

Performing segmentation and fracture grade classification in a multi-task manner can address these limitations effectively. A multi-task learning framework enables the model to leverage shared representations between tasks, enhancing the overall performance. For instance, the features learned during the segmentation process, such as detailed vertebral morphology, can be directly applicable to the identification and classification of fracture grades. This shared learning approach improves the efficiency and accuracy of the model by providing a richer context for each task.

Multi-task learning has been proven to be successful in medical image analysis [41] [42], as it can improve learning efficiency and performance by leveraging the inductive bias when jointly solving related tasks [43]. By integrating vertebral segmentation and fracture classification into a unified framework, we allow the model to benefit from the detailed anatomical features captured during segmentation, leading to more accurate fracture detection and grade classification. This holistic approach has potential to over-

come the limitations of previous methods and deliver more precise and comprehensive diagnostic tools for vertebral fractures in CT scans.

Instead of requiring separate models for vertebrae segmentation and fracture classification, a multi-task model can share common features and representations, making the process more data efficient and multi-task learning has been shown to provide faster learning speed for related tasks, helping to alleviate the weaknesses of deep learning models: large-scale data requirements and computational demand [44]. This integration can also facilitate better generalisation to diverse datasets [45], as the model learns to handle variations in vertebral anatomy and fracture presentations cohesively.

2.5 Project Objectives and Contributions

The primary objective of this project is to develop a two-stage pipeline with an integrated multi-task learning framework in the latter stage that combines 3D vertebrae segmentation and fracture grade classification within a single model. This involves several specific tasks aimed at improving the current state of automated vertebral fracture detection. The first stage seeks to achieve accurate localisation of vertebrae in CT scans. This initial step is crucial because the vertebrae centers must be precisely located before they can be segmented and assessed for fractures. We emphasise that without accurate localisation, the subsequent segmentation and classification tasks would be unreliable and prone to significant errors. Following individual vertebrae localisation, precise 3D segmentation of individual vertebrae will be performed to provide detailed anatomical delineations, which are essential for accurate fracture detection. We concurrently perform classification of vertebral fracture grades by leveraging shared representations between segmentation and classification tasks in a multi-task manner. This integrated approach is designed to create an efficient workflow that reduces computational costs, improves data efficiency, and enhances the diagnostic accuracy of vertebral fractures. Ultimately, this project seeks to improve clinical decision-making and patient care by providing a robust and precise diagnostic tool for both targeted and incidental detection of vertebral fractures.

Chapter 3

Methods

3.1 Overview

This project employs a two-stage pipeline for automated vertebral fracture detection and grading in 3D CT scans. The primary objective is to develop a robust and efficient system that leverages multi-task learning to enhance diagnostic accuracy.

In the first stage, our pipeline focuses on the precise localisation of individual vertebrae within CT images and we frame this task as a landmark detection problem. We employ a U-Net-based network architecture to generate heatmaps and offset maps. The heatmaps provide a probabilistic representation of vertebrae locations, while the offset maps indicate the displacement vectors from each voxel to the nearest vertebra center. This combination allows for the accurate identification of vertebrae centers within the CT scans. The localisation network includes multiple convolutional blocks, max pooling layers, and transposed convolutional layers to progressively capture and refine spatial features. The accurate localisation of vertebrae centers is crucial as it isolates the regions of interest for subsequent analysis.

In the second stage, the pipeline processes these identified vertebrae centers for detailed analysis. For each identified vertebra center, a cropped patch is extracted from the original CT image, ensuring the target vertebra is fully enclosed. These patches are then fed into a multi-task network designed to perform both 3D segmentation and fracture classification simultaneously. The multi-task network uses a shared encoder-decoder U-Net architecture similar to the localisation network but with two output branches, one for segmentation and another for classification.

3.2 Vertebrae Localisation

The initial stage of our pipeline involves accurately localising the individual vertebrae within arbitrary field-of-view (FOV) CT images, where the number of vertebrae present in each image may vary. To do so, we frame it as a landmark detection task, where the landmarks themselves are the vertebrae centers. The intuitive solution would be to directly regress the vertebrae center coordinates given any arbitrary FOV CT image [46] [47]. However, while computationally efficient, this method has been shown to be inaccurate or to miss landmarks altogether [48]. Subsequently, as previous work has performed, we could perform heatmap regression by fitting a 3D Gaussian kernel at each vertebra center, generating a heatmap where the intensity values represent the proximity to these centers. Regression is then performed on the heatmap values [49] [50]. However, pure heatmap regression is prone to failure in CT scans with a large number of vertebrae, as these vertebrae are more tightly packed compared to scans with fewer vertebrae [34]. Given the severe consequences of undiagnosed vertebral fractures and, by extension, the critical importance of precise vertebrae localisation, regardless of the number of vertebrae present, we seek a more robust method for accurately localising vertebrae centers. Figure 3.1 provides an overview of our localisation network, which is discussed in detail below.

3.2.1 Localisation Network Model Architecture

Our vertebrae localisation network largely follows the implementation by Cui et al. [34] and is based on a standard U-Net architecture [36]. Structured in an encoder-decoder format, the model features multiple layers of convolutional blocks, max pooling, and transposed convolutional layers. Each convolutional block consists of two 3D convolutional layers, followed by instance normalisation [51] and a LeakyReLU activation function [52]. Note: all components and methods were implemented from scratch.

The encoder path consists of five convolutional blocks, each followed by a 3D max pooling layer, which progressively reduces the spatial dimensions while increasing the number of feature channels. The decoder path reverses this process by using transposed convolutional layers [53] to progressively upsample the feature maps, thereby restoring the original spatial dimensions of the input. Each upsampling step is followed by a concatenation with the corresponding encoder feature maps, which is then processed through additional convolutional blocks. This skip-connection ensures that the decoder has access to high-resolution features from the encoder, enhancing the network's ability

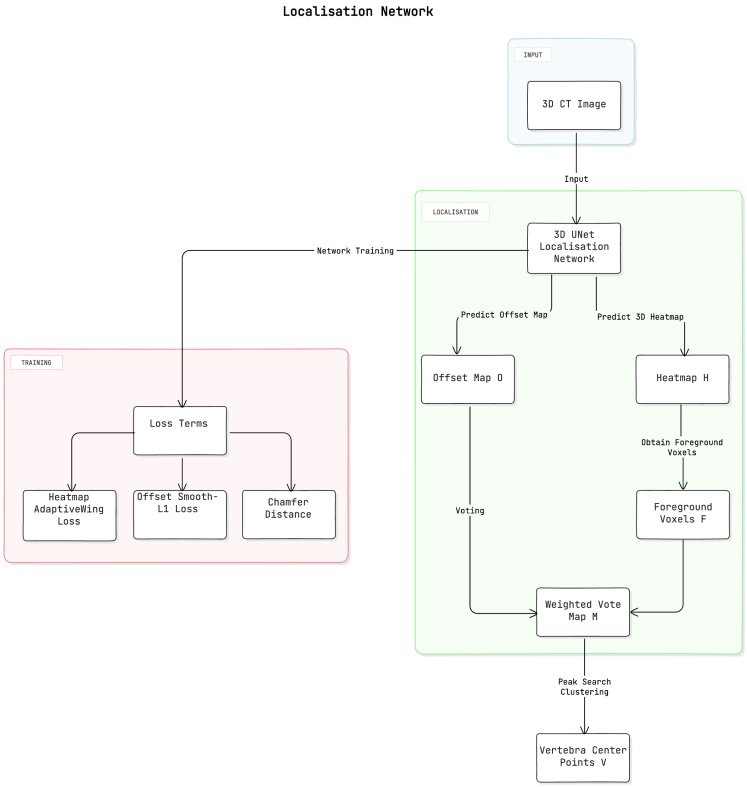

Figure 3.1: Localisation network to predict vertebrae centers given an input CT image

to reconstruct detailed spatial information [36].

The localisation network features two output branches and takes as input a 3D CT image. It predicts a one-channel Gaussian 3D heatmap (H) and a three-channel offset map (O), where the spatial dimensions of both are the same as the input CT image. The heatmaps are derived from the ground-truth vertebrae center coordinates by fitting a 3D Gaussian with a standard deviation of $\delta = 3$ voxel-size, providing a probabilistic map of vertebrae locations. In order to output valid probabilities bounded between 0 and 1, the output layer of the heatmap branch uses a sigmoid activation.

To generate the offset map from the ground-truth center coordinates, a meshgrid for the coordinates is created over the entire 3D volume. For each vertebra center, the Euclidean distance from the center to all voxels is computed, and the offset map is updated with the relative offset vectors for voxels where this center is the nearest resulting in a three-channel map which indicates the 3D offset (displacement) vectors of each voxel pointing to its nearest vertebra center.

3.2.2 Weighted Vote Map Generation and Vertebrae Center Localisation

To perform localisation, we obtain all foreground voxels, V, from the predicted heatmaps, $\hat{H}$, by thresholding voxels above the value of 0.8 ($\hat{H} > H_t = 0.8$). For each foreground voxel $v_i \in V$, we take its corresponding 3D offset vector from the predicted offset map, $\hat{O}$ as the vote from v_i to its nearest vertebra center and take the heatmap value of v_i as the weight of that vote. For instance, a foreground voxel at position $[0,0,1]$ in $\hat{H}$ with a value of 0.8 and a corresponding offset vector of $[0,1,1]$ would contribute to the voxel at position $[0,1,2]$, calculated as $[0,0,1]+[0,1,1]$, with a weight of 0.8. The weighted votes are subsequently accumulated in a weighted vote map, $\hat{M}$, that takes the same spatial dimension as $\hat{H}$, as outlined in Algorithm 1 below.

Algorithm 1 Weighted Vote Map Generation

Input: Predicted heatmap $\hat{H}$, Predicted offset map $\hat{O}$, Heatmap threshold H_t.
Output: Weighted vote map $\hat{M}$.
$V \leftarrow \{v_i \mid \hat{H}(v_i) > H_t\}$ ▷ Find foreground voxels
if V is empty **then**
    Continue to next image ▷ No valid votes, skip image
end if
$M \leftarrow 0$ ▷ Initialise vote map to zero
for each voxel $v_i \in V$ **do**
    $(x,y,z) \leftarrow$ coordinates of v_i
    $h \leftarrow \hat{H}(x,y,z)$ ▷ Vote value from predicted heatmap
    $(\Delta x, \Delta y, \Delta z) \leftarrow \hat{O}(x,y,z)$ ▷ Extract predicted offsets
    $(vx, vy, vz) \leftarrow (x+\Delta x, y+\Delta y, z+\Delta z)$ ▷ Compute vote locations
    if $0 \leq vx < \dim_x(M)$ and $0 \leq vy < \dim_y(M)$ and $0 \leq vz < \dim_z(M)$ **then**
        $\hat{M}(vx, vy, vz) \leftarrow \hat{M}(vx, vy, vz) + h$ ▷ Accumulate vote value
    end if
end for
Return: M

As a post-processing step, we perform clustering to localise the density peaks in the weighted vote map, $\hat{M}$, which correspond to the coordinates of the vertebrae centers. To that end, we implement a variant of the fast-search clustering algorithm introduced by Rodriguez et al. [54]. Specifically, the algorithm identifies density

peaks in $\hat{M}$ by iterating through each voxel and comparing it to an empirically chosen neighborhood of $3 \times 3 \times 3$ voxels. For each voxel $(x, y, z) \in \hat{M}$, if the voxel's value is greater than a value threshold, $\eta = 0.1$, and is also the maximum value within its neighborhood, it is considered a peak. Additionally, the algorithm checks the distance to higher density voxels in $\hat{M}$ to ensure significant separation, confirming a peak only if this distance exceeds a distance threshold, λ. This ensures that the identified centers are well-separated and represent significant density maxima. We refer the reader to Algorithm 2 in Appendix A.1 for the full pseudocode implementation for the fast-search clustering algorithm.

3.2.3 Model Training

3.2.3.1 Loss Functions

To train the localisation network, we utilise several loss terms. For the one-channel heatmap regression that outputs $\hat{H}$, pixel-wise $L1$ or $L2$ Loss would be natural choices [55] [56] [57]. However, these loss functions have been shown to have performance limitations: they are not sensitive to small errors, which hinders the robust localisation of the Gaussian kernel's mode. Furthermore, they treat all voxels equally, causing background voxels (which tend to be in the large majority) to dominate the Loss. Consequently, models trained with pixel-wise $L1$ or $L2$ Loss tend to predict blurry heatmaps with low intensity on foreground voxels relative to the ground truth, leading to inaccurate landmark localisation [58].

In the work of Cui et al. [34], Smooth-$L1$ Loss was employed for training the heatmaps. However, we observed that this approach led to unstable training where background voxels were disproportionately influencing the loss, resulting in all voxel values being driven towards zero. Consequently, no voxels would exceed the threshold H_t and as such, the offset map was unable to train. To address this limitation, we adopted the Adaptive Wing Loss function as proposed by Wang et al. [58] for heatmap regression. This loss function is designed to adapt its curvature based on the values of the ground truth voxels. As training progresses, the influence on foreground voxels increases as the errors decrease, focusing more on reducing these errors. Conversely, this influence rapidly decreases as errors approach zero, thus preventing overfitting. For background voxels, the influence of the loss function gradually tends to zero as errors decrease, reducing the focus on these voxels and stabilising the training process.

$$\mathcal{L}_{\text{AdaptiveWingLoss}} = \begin{cases} \omega \ln\left(1 + \left|\frac{y-\hat{y}}{\varepsilon}\right|^{\alpha-y}\right) & \text{if } |y-\hat{y}| < \theta \\ A|y-\hat{y}| - C & \text{otherwise} \end{cases}$$

where y and $\hat{y}$ are the ground truth and predicted heatmap voxel values, respectively. $\omega, \theta, \varepsilon$ and α are positive values, $A = \omega\left(1/\left(1+(\theta/\varepsilon)^{\alpha-y}\right)\right)\left((\alpha-y)(\theta/\varepsilon)^{(\alpha-y-1)}\right)(1/\varepsilon)$ and $C = (\theta A - \omega \ln(1+(\theta/\varepsilon)^{\alpha-y}))$ are used to make the loss function continuous and smooth at $|y-\hat{y}| = \theta$. Through experimentation, we find that setting $\omega = 10$, $\theta = 0.5$, $\varepsilon = 3$, $\alpha = 2.1$ resulted in the most stable training of the heatmaps.

For the three-channel offset map, $\hat{O}$, we supervise its training using the Smooth-$L1$ Loss. However, we recall that $\hat{H}$ and $\hat{O}$ are trained simultaneously, but the voxels in $\hat{H}$ only take values in the range $[0,1]$, while the voxels in $\hat{O}$ can theoretically take values in the range $[-D_x, D_x]$, $[-D_y, D_y]$, and $[-D_z, D_z]$ for the x-, y-, and z-offsets, respectively, where D_x, D_y, and D_z are the dimensions of the input CT image minus one. As such, to ensure that the loss from background voxels $\hat{O}$ does not dominate and negatively impair model gradients, we mask the Smooth-$L1$ Loss computation of $\hat{O}$ to include only the foreground voxels identified in $\hat{H}$. This results in the following formulation for our masked Smooth-$L1$ Loss:

$$M = \mathbb{1}\{\hat{H} > H_t\}$$

$$\mathcal{L}_{\text{Smooth}L1} = \frac{1}{N} \sum_{i \in \{x,y,z\}} \sum_{j,k,l} M_{jkl} \cdot \text{Smooth}L1(\hat{O}_{ijkl} - O_{ijkl})$$

$$\text{Smooth}L1(d) = \begin{cases} \frac{0.5d^2}{\beta} & \text{if } |d| < \beta \\ |d| - 0.5\beta & \text{otherwise} \end{cases}$$

M is an indicator function that equals 1 if $\hat{H}$ is greater than the threshold H_t, and 0 otherwise. d is the difference $\hat{O}_{ijkl} - O_{ijkl}$ and β is a positive value. This criterion uses a squared term if the absolute element-wise error $|d|$ falls below β, and an $L1$ term $|d| - 0.5\beta$ otherwise. We use the default value of $\beta = 1$.

To further robustly regress the vertebrae centers, we derive a candidate vertebrae center set, $\hat{C}$, which is attained by thresholding the generated weighted vote map, M, to extract its foreground voxels:

$$\hat{C} = \{(x,y,z) \mid M(x,y,z) > V_t\},$$

where V_t is the vote threshold parameter. From our experiments, the choice of $V_t = 0.6$ resulted in the most stable training, as it produced a balanced number of selected

candidate centers and resulted in the most accurate localisation on a held-out validation set. Subsequently, we compute the Chamfer Distance between all candidate centers in $\hat{C}$ and the ground-truth center coordinates, C, in order to minimise the bidirectional distance between points in the two:

$$\mathcal{L}_{\text{CD}} = \sum_{\hat{c}_i \in \hat{C}} \min_{c_k \in C} \|\hat{c}_i - c_k\|_2^2 + \sum_{c_k \in C} \min_{\hat{c}_i \in \hat{C}} \|c_k - \hat{c}_i\|_2^2 .$$

As such, the total training loss, $\mathcal{L}_{\text{Localisation}}$, is formulated as:

$$\mathcal{L}_{\text{Localisation}} = \mathcal{L}_{\text{AdaptiveWingLoss}} + \phi \mathcal{L}_{\text{SmoothL1}} + \gamma \mathcal{L}_{\text{CD}}.$$

ϕ and γ are balancing weights for the loss such that no one loss dominates training and are empirically set to $\alpha = 0.01$ and $\beta = 0.01$.

3.2.3.2 Dataset

Throughout the development and evaluation of the project pipeline, we utilise the original, publicly available VerSe2019 dataset[1]. Developed for the vertebral labeling and segmentation challenge at the MICCAI 2019 conference, this dataset comprises 141 patients with 160 3D CT images, with some patients having multiple scans [59]. Furthermore, the dataset is pre-divided into training (n=80), validation (n=40), and test (n=40) sets.

The data, acquired from multiple CT scanners across various sites, includes a range of fields of view (cervical, thoraco-lumbar, and cervico-thoraco-lumbar scans) [59]. Consequently, the number of visible vertebrae varies across scans and furthermore, the images may differ in orientation and spacing. The VerSe2019 dataset additionally includes radiologist-refined 3D segmentation masks, aligned in orientation and spacing with their corresponding images. It also provides ground-truth vertebrae center coordinates in an accompanying JSON file. Notably, the dataset provides ground-truth fracture grades derived from the Genant semi-quantitative method for all thoracic and lumbar vertebrae in each CT image, but not for the cervical vertebrae. Therefore, while our localisation network addresses all three spinal regions, the fracture detection stage and its associated results and discussion focus exclusively on thoracic and lumbar vertebrae.

[1]The choice of VerSe2019 over VerSe2020 is due to the availability of ground-truth fracture data for VerSe2019.

3.2.3.3 Localisation Network Implementation Details

As a preprocessing step, we resample all images to 1mm $\times$ 1mm $\times$ 1mm, using trilinear interpolation, to ensure consistent voxel dimensions across all spatial axes. As input to the localisation network, all CT images are randomly cropped to $128 \times 128 \times 128$. While larger crop sizes were considered, this choice was limited by computational constraints. As a result of the large variance with regard to the size of the images in our dataset (min = $103 \times 157 \times 76$ vs max = $915 \times 1189 \times 709$), the chosen crop size does not necessarily guarantee a CT crop with a valid vertebra center in it (full statistics on dataset size are in Appendix A.2). As such, to ensure at least one valid vertebra center is enclosed, all images that exceed $256 \times 256 \times 256$ are first downsampled to that size prior to random cropping. While there is no definitive consensus on the most optimal interpolation technique for image downsampling [60], trilinear interpolation has been selected for our purposes due to its computational efficiency and demonstrated efficacy in minimising interpolation error with medical images [61]. Conversely, for images smaller than the crop size of $128 \times 128 \times 128$, we perform zero padding of the image.

Finally, for model training, we employ He initialisation [62] for the model weights and utilise the Adam [63] optimiser with a learning rate of 0.05, training the network for 1000 epochs. Early stopping is not implemented in this process because the heatmaps take time to train. With Adaptive Wing Loss, all voxel values initially decrease towards zero before any foreground voxel values start to rise above the heatmap threshold H_t. Until the point where any voxels exceed this threshold, the offset and Chamfer Distance losses are set to zero. Stopping the training early would prevent the model from utilising these loss functions once the threshold is surpassed. As such, we implement regular saving of the model and monitor the loss curves. At inference time, we use the model saved at the epoch where training loss plateaus, which was at epoch 600.

3.3 Vertebrae Fracture Detection

Following the localisation of vertebrae centers, we proceed to detect vertebral fractures and their grades. To prepare a suitable training dataset for fracture detection, we first extract $96 \times 96 \times 96$ crops centered on the ground-truth vertebra centers from each CT image in the VerSe2019 training set, which has been isometrically resampled to 1mm $\times$ 1mm $\times$ 1mm using trilinear interpolation. For the specific vertebra of interest, we extract its segmentation mask from the original CT segmentation file, which has also

been resampled to isotropic spacing using nearest neighbour interpolation. This results in a new dataset of individual vertebrae crops and their accompanying segmentation mask, each labeled with its ground-truth fracture grade, which is subsequently used to train the fracture detection network.

3.3.1 Fracture Detection Network Model Architecture

As elucidated earlier, we aim to perform segmentation and fracture classification together in a multi-task manner and, in doing so, compare and evaluate the performance benefits of using shared representations for fracture grade classification. To this end, we implement two separate fracture detection networks: ***FracNet***, which solely performs fracture grade classification given the aforementioned cropped CT patch ($\hat{y} = f(x)$), where $\hat{y}$ represents the predicted fracture grade and x represents the input CT patch), and ***FracSegNet***, which performs simultaneous 3D segmentation of the target vertebra and fracture grade classification. During training, ***FracSegNet*** receives both the cropped CT patch and the corresponding vertebra ground-truth segmentation mask ($\{x, S\}$), where S is the ground-truth segmentation mask, and the model is trained to predict the fracture grade and segmentation mask ($\{\hat{y}, \hat{S}\} = f(x, S)$). During inference, ***FracSegNet*** only receives the cropped CT patch as input ($\{\hat{y}, \hat{S}\} = f(x)$), and the model outputs both the predicted fracture grade $\hat{y}$ and the predicted segmentation mask $\hat{S}$.

The backbone of both networks is architecturally identical to that of the localisation network but differs in terms of the output branches. Following the final decoder layer, ***FracNet*** employs an adaptive global average pooling layer [64], which reduces the spatial dimensions to a single vector per feature map. This is followed by a fully connected (FC) layer with 256 neurons, after which a dropout layer is applied to mitigate overfitting. The next FC layer has 128 neurons, followed by another dropout layer. Subsequently, an FC layer with 64 neurons is applied, followed by a final fully connected layer with 4 neurons corresponding to the number of fracture grades $\{0, 1, 2, 3\}$. Barring the final FC layer, all others are followed by a ReLU activation function. Finally, a softmax activation function is used on the output of the final FC layer to output a valid probability distribution across grades.

In contrast, ***FracSegNet*** retains the classification branch described above and adds a binary segmentation branch that applies a $1 \times 1 \times 1$ convolutional layer to the output of the final decoder layer, generating a feature map that matches the spatial dimensions of the input. This is followed by a sigmoid activation function to produce the final 3D

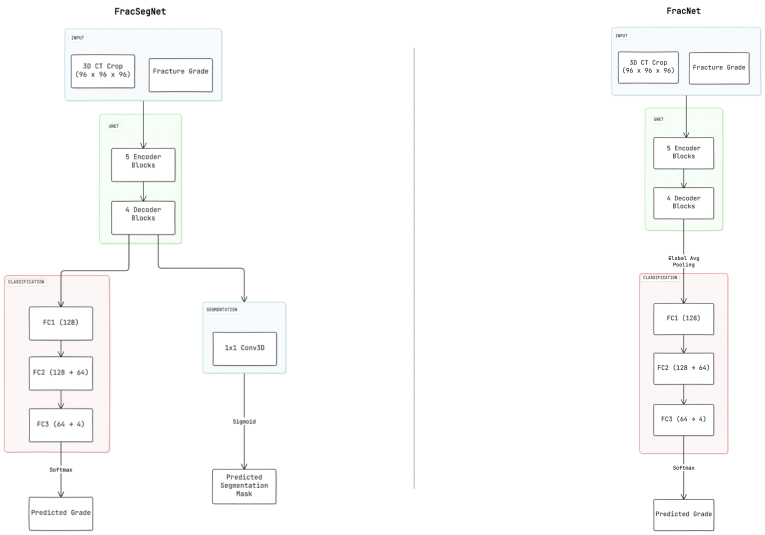

Figure 3.2: Visual schematic of the architectures for FracNet and FracSegNet. Both models share an identical UNet backbone for feature extraction. FracNet employs a classification branch with global average pooling and fully connected layers for fracture grade classification. FracSegNet extends this architecture by adding a parallel segmentation branch, which applies a 1x1x1 convolution followed by a sigmoid activation to generate a 3D segmentation mask of the vertebra.

segmentation mask. Figure 3.2 provides a visual overview of both the ***FracNet*** and ***FracSegNet*** architecture.

3.3.2 Model Training

3.3.2.1 Loss Functions

We supervise the training of ***FracNet*** using Focal Loss as introduced by Lin et al. [65] for the task of dense object detection. We recall that the VerSe2019 dataset presents a significant class imbalance concerning fracture grades, with the new fracture detection training set comprising 591 Grade 0, 63 Grade 1, 50 Grade 2, and 22 Grade 3 crops. Addressing this imbalance is required to prevent the model from being biased toward predicting Grade 0. In doing so, Focal Loss introduces a scaling factor, γ, to the standard Cross-Entropy Loss that down-weights the contribution of easy-to-classify examples, allowing the model to focus more on hard, misclassified cases. Such an adjustment aids in improving the model's performance on minority classes by emphasising learning from difficult examples, which are underrepresented in imbalanced datasets like ours.

Similar to the implementation by Lin et al., we further address the class imbalance in

the dataset by computing the weighting factor α_c for each class c based on the inverse of the class frequencies ([591, 63, 50, 22]). Specifically, α_c is calculated as the normalised inverse of the number of samples in each class, ensuring that classes with fewer samples are given higher importance during training.

The Focal Loss is thus defined as:

$$\mathcal{L}_{\text{Focal}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \alpha_c (1-\hat{y}_{ic})^{\gamma} y_{ic} \log(\hat{y}_{ic}) \tag{3.1}$$

where α_c is a weighting factor for class c computed from the inverse of class counts to handle class imbalance, γ is the scaling parameter that adjusts the rate at which examples with more confident predictions (higher $\hat{y}_{ic}$) are down-weighted, $C = 4$ is the number of classes, y_{ic} is a binary indicator that is 1 if the i-th sample belongs to class c and 0 otherwise, and $\hat{y}_{ic}$ is the predicted probability for the i-th sample belonging to class c. We follow the implementation in Lin et al. by setting $\gamma = 2$.

The classification branch of ***FracSegNet*** is supervised with the same formulation of Focal Loss as above. For the segmentation branch, we use the Dice Loss which is a widely used measure of overlap between predicted and ground truth segmentations [66]. Note: Dice Loss is equivalent to $1 -$ Dice Score. The binary variant of Dice Loss is shown below:

$$\mathcal{L}_{\text{DiceLoss}} = 1 - \underbrace{\frac{\sum_{n=1}^{N} p_n r_n + \varepsilon}{\sum_{n=1}^{N} p_n + r_n + \varepsilon} - \frac{\sum_{n=1}^{N} (1-p_n)(1-r_n) + \varepsilon}{\sum_{n=1}^{N} 2 - p_n - r_n + \varepsilon}}_{\text{Dice Score}},$$

where the ε term is to avoid division by 0. Finally, our total training loss for FracSegNet, $\mathcal{L}_{\text{FracSegNet}}$ is formulated as:

$$\mathcal{L}_{\text{FracSegNet}} = \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{DiceLoss}},$$

where $\mathcal{L}_{\text{Focal}}$ and $\mathcal{L}_{\text{DiceLoss}}$ are equally weighted.

3.3.2.2 Fracture Detection Network Implementation Details

In terms of model training, we use He initialisation for both ***FracNet*** and ***FracSegNet*** and train both models with the Adam optimiser and a learning rate of 0.0005 for 1000 epochs with an early stopping patience of 50 epochs. We adjust the learning rate dynamically as well, reducing the learning rate by a factor of 10 every 80 steps.

Chapter 4

Experiments

4.1 Vertebrae Localisation

To quantitatively evaluate our localisation network on the test set, we begin by randomly cropping a $128 \times 128 \times 128$ patch from each testing image, similar to the training process. From these patches, we generate the predicted heatmaps, $\hat{H}$, and offset maps, $\hat{O}$, which are used to compute the weighted vote map, M, as detailed in Algorithm 1. We then apply fast-search clustering (described in Algorithm 2) on M to localise the predicted centers. The *localisation error* is then measured as the $L2$ norm (Euclidean distance) between each predicted vertebra center and its nearest ground-truth center. Furthermore, the mean localisation error between predicted and ground-truth centers are calculated for each vertebra region—cervical, thoracic, and lumbar—allowing for region-specific analysis. For comparison, we highlight the performances of 2 localisation methods: the full localisation network as described, ***LocNet-F***, and a heatmap only localisation network, ***LocNet-H***, where fast-search clustering is performed directly on $\hat{H}$. As mentioned in 3.2.2, for the fast-search clustering in ***LocNet-F***, we set $\eta = 0.1$ and $H_t = 0.7$ as it resulted in the most accurate localisation of vertebrae centers on the validation set. For the same reason, we set $\eta = 0.7$ for ***LocNet-H***. For both, we set $\lambda = 5.0$.

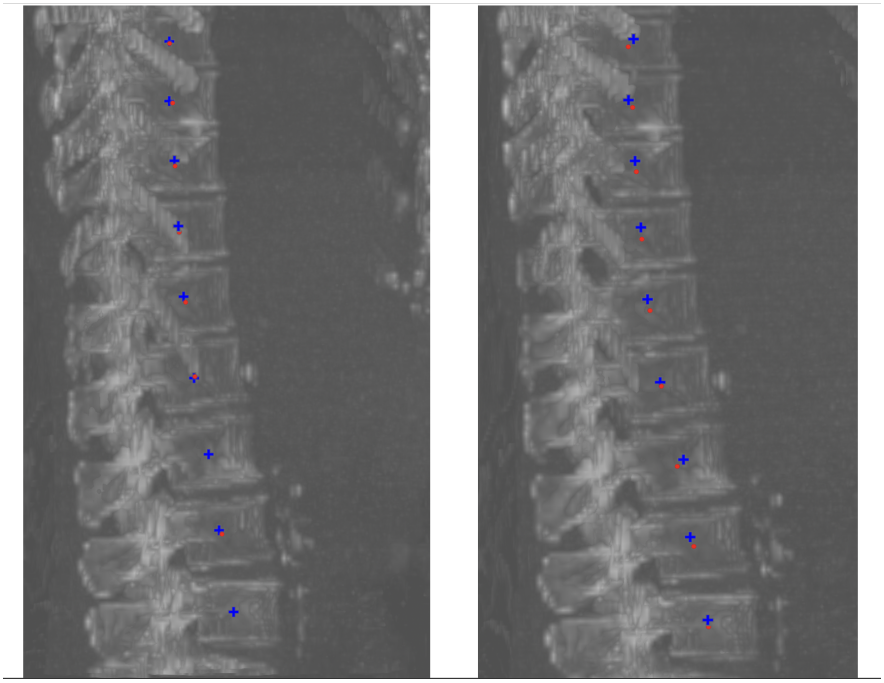

Figure 4.1: Example of localisation results for a single input CT for ***LocNet-H*** (left) and ***LocNet-F*** (right). Blue denotes the predicted centers while red is ground-truth.

Method	Localisation Error $(mm)\downarrow$			
	Cervical	Thoracic	Lumbar	Overall
LocNet-H	1.19 $\pm$0.5	3.61 $\pm$8.2	6.06 $\pm$14.6	4.27 $\pm$10.9
LocNet-F	2.58 $\pm$2.1	3.78 $\pm$6.2	3.09 $\pm$5.4	3.39 $\pm$5.6

Table 4.1: Quantitative vertebrae localisation results for the full localisation network, ***LocNet-F***, and the heatmap only network, ***LocNet-H***.

4.2 Vertebrae Fracture Detection

To assess the performance of our vertebrae fracture detection network, it would have been ideal to take vertebrae crops based on the predicted vertebrae centers provided by ***LocNet-F*** for a given input CT image, as its predictions have demonstrated robust localisation accuracy. However, to ensure a precise and consistent evaluation, we utilise the ground truth vertebrae centers. This approach allows us to eliminate any potential biases introduced by prediction errors from ***LocNet-F***, thereby providing a more accurate measure of the fracture detection network's capabilities on its own. Thus, to create a standardised test set to evaluate ***FracNet*** and ***FracSegNet***, we follow the

same procedure as for the training set and take $96 \times 96 \times 96$ sized crops of every vertebra present across all CT images in the Verse2019 test set. It is important to note that the test set for ***FracNet*** is larger than that for ***FracSegNet***, as there are fewer available segmentation masks than vertebra centroids. This discrepancy arises because some vertebrae, particularly those at the peripheries of certain CT images lack corresponding segmentation masks.

4.2.1 Classification Metrics

In evaluating the performance of the models for fracture grade prediction, the following classification metrics were employed: **precision**, **recall**, and **F1-score**. These metrics were chosen to provide a comprehensive assessment of each model's capability, particularly in the context of the inherent class imbalance present in our task of grading vertebral fractures, where grade 2 and grade 3 fractures occur far less frequently. **Accuracy** was deliberately omitted from our evaluation due to its potential to be misleading in such imbalanced scenarios, where it may disproportionately reflect performance on the majority class. Furthermore, we report precision, recall, and F1-score individually for each grade, to ensure that the performance on underrepresented, and more clinically important, fracture grades is properly accounted for despite the imbalance in the dataset.

Precision is critical for evaluating the accuracy of the model's positive predictions for each fracture grade. It is defined as the ratio of true positives to the sum of true positives and false positives, reflecting the proportion of correct positive predictions among all cases predicted as positive for a specific grade. In a clinical setting, where the consequences of misclassification can be significant, the precision of such a diagnostic tool for classification is crucial.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall, also known as sensitivity, measures the model's ability to identify all true instances of a particular fracture grade. It is the ratio of true positives to the sum of true positives and false negatives. Recall is crucial in the context of fracture detection, as it ensures that the model is capable of identifying all true cases of a given fracture grade, thereby reducing the risk of missing a severe fracture. As aforementioned, the implications of missed fracture diagnoses are stark, which informed the selection of recall as a metric in evaluating model performance.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score represents the harmonic mean of precision and recall, providing a single

measure that balances the these two metrics. It is particularly useful when evaluating models on imbalanced datasets [67], as it evaluates the models performance not only on the majority class but also on minority classes, which are of greater clinical relevance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.2.2 Segmentation Metrics

To evaluate the performance of the models for vertebrae segmentation by ***FracSegNet***, we employ the following segmentation metrics: **accuracy**, **precision**, **recall**, **Dice coefficient**, and **average Hausdorff distance**.

Accuracy measures the overall correctness of the segmentation by calculating the proportion of correctly predicted pixels (both true positives and true negatives) out of the total number of pixels. However, in the context of imbalanced datasets—where background pixels vastly outnumber foreground pixels, comparative to other images in the dataset—accuracy can be misleading, as it disproportionately reflects performance on the majority class (background). Therefore, while included for completeness, accuracy must be considered alongside the following metrics that better reflect ***FracSegNet's*** performance on the foreground class.

Precision in segmentation evaluates the accuracy of the positive (foreground) predictions by calculating the ratio of true positive voxels to the sum of true positive and false positive voxels. This metric determines the proportion of correctly identified vertebrae regions among all regions predicted as vertebrae, which is particularly important in avoiding over-segmentation.

Recall (also known as sensitivity) measures the model's ability to capture all true positive voxels, defined as the ratio of true positive voxels to the sum of true positive and false negative voxels. High recall indicates that the model successfully identifies most of the vertebral regions, which is essential in a clinical setting to avoid missing any regions of the vertebrae. The formulation for accuracy, recall and precision are identical to that of the classification task, but apply voxel-wise in the segmentation context.

Used earlier in the Dice Loss, **Dice Score** is a widely used metric in medical image segmentation that evaluates the overlap between the predicted and ground truth segmentation masks. We refer the reader to the aforementioned formulation of Dice Loss in Section 3.3.2.1 for the formulation of Dice Score.

Average Hausdorff Distance is a metric that quantifies the spatial distance between the predicted and ground truth segmentation boundaries. It is defined as the average

of the minimum distances between each point in one set and the nearest point in the other set, calculated in both directions (from predicted to ground truth and vice versa). Formally, for two finite point sets X and Y, the average Hausdorff distance is given by [68]:

$$d_{\mathrm{AHD}}(X,Y) = \left(\frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} d(x,y) + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} d(x,y) \right) /2$$

Where:

- $|X|$ and $|Y|$ are the number of points in sets X and Y, respectively.
- $d(x,y)$ represents the Euclidean distance between points x and y.
- The summations calculate the average minimum distance from each point in X to the closest point in Y, and vice versa.

4.2.3 Results

In our experiments, we observed suboptimal training performance for the classification task in both ***FracNet*** and ***FracSegNet***, as illustrated by the training curves in Figure 4.2. Despite the implementation of Focal Loss, ***FracNet*** exhibited a steady decline in training loss; however, the validation loss remained stagnant from the very onset of training, indicating a potential failure to generalise. In the case of ***FracSegNet***, neither the training nor validation losses showed significant improvement throughout the training process, despite the segmentation task seeing stable training. Additionally, it was observed that at the onset of training, the scales of the Focal Loss for classification and the Dice Loss for segmentation varied considerably. As a result, weighting schemes were attempted to balance these losses, but these efforts did not yield any improvement.

To that end, we experiment with weighted and unweighted Cross-Entropy Loss as the loss function for classification:

$$L_{\mathrm{CE}}(p,y) = -\sum_{c=1}^{4} y_c \log(p_c)$$

$$L_{\mathrm{WCE}}(p,y) = -\sum_{c=1}^{4} w_c \cdot y_c \log(p_c)$$

where:

- $p = [p_1, p_2, p_3, p_4]$ represents the predicted probability distribution over the 4 classes for a given input crop.

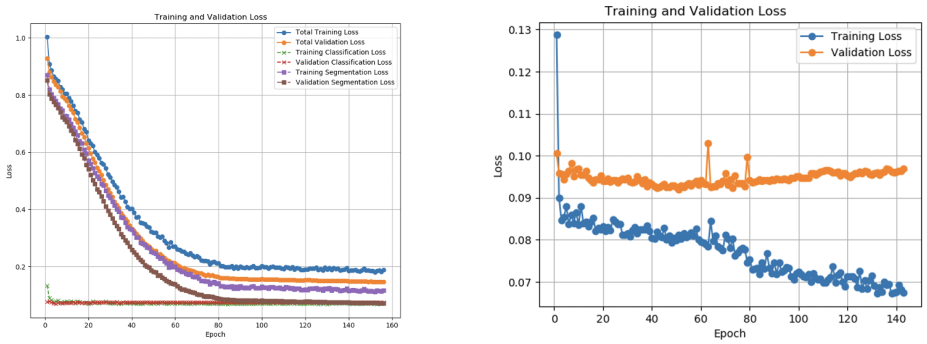

Figure 4.2: Training and validation loss curves for ***FracSegNet*** (left) and ***FracNet*** (right) with Focal Loss for classification.

- $y = [y_1, y_2, y_3, y_4]$ is the one-hot encoded true class label, where $y_c = 1$ if the class label is c, and $y_c = 0$ otherwise.
- $w_c = \frac{1/n_c}{\sum_{c'=1}^{4}(1/n_{c'})}$ is the weight assigned to class c, calculated as the normalised inverse of the class count n_c, where n_c is the number of samples in class c.

In our experiments, some degree of stable training was observed for ***FracSegNet*** using unweighted Cross-Entropy Loss, while ***FracNet*** exhibited a failure to generalise with both weighted and unweighted Cross-Entropy Loss, similar to its training dynamics with Focal Loss. As such we include the results for ***FracSegNet*** with unweighted Cross-Entropy Loss alongside both ***FracNet*** and ***FracSegNet*** that were trained with Focal Loss in Table 4.3 which shows the confusion matrix of ground-truth fracture grades versus predicted fracture grades for all 3 networks. Accordingly, we highlight the precision, recall and F1-score for the predictions of each grade for all 3 networks in Table 4.4.

		Accuracy	Precision	Recall	Dice Score	Avg. Hausdorff Distance
FracSegNet (Focal Loss)	Grade 0	0.98	0.73	0.76	0.73	16.7
	Grade 1	0.98	0.73	0.76	0.73	15.0
	Grade 2	0.97	0.73	0.67	0.68	17.61
	Grade 3	0.97	0.66	0.71	0.67	16.9
FracSegNet (Unweighted CE)	Grade 0	0.99	0.92	0.89	0.90	12.53
	Grade 1	0.99	0.91	0.91	0.91	11.46
	Grade 2	0.99	0.89	0.93	0.91	12.99
	Grade 3	0.98	0.84	0.90	0.87	13.41

Table 4.2: Comparison of segmentation performance for FracSegNet (Focal Loss) and FracSegNet (Unweighted CE) across different fracture grades.

			Predicted Grade				
			Grade 0	Grade 1	Grade 2	Grade 3	Support
FracNet (Focal Loss)	GT Grade	Grade 0	292	0	0	0	292
		Grade 1	38	0	0	0	38
		Grade 2	22	0	0	0	22
		Grade 3	12	0	0	0	12
FracSegNet (Focal Loss)		Grade 0	257	0	0	0	257
		Grade 1	36	0	0	0	36
		Grade 2	22	0	0	0	22
		Grade 3	12	0	0	0	12
FracSegNet (Unweighted CE)		Grade 0	235	20	2	0	257
		Grade 1	31	3	2	0	36
		Grade 2	8	6	7	1	22
		Grade 3	3	2	7	0	12

Table 4.3: Confusion matrix comparing the predicted fracture grades against the ground truth (GT) grades for both FracNet (Focal Loss) and both FracSegNet networks. The table shows the number of vertebrae that were predicted as each grade (columns) for each ground truth grade (rows).

		Precision	Recall	F1-score
FracNet (Focal Loss)	Grade 0	0.80	1.00	0.89
	Grade 1	-	0.00	0.00
	Grade 2	-	0.00	0.00
	Grade 3	-	0.00	0.00
FracSegNet (Focal Loss)	Grade 0	0.79	1.00	0.88
	Grade 1	-	0.00	0.00
	Grade 2	-	0.00	0.00
	Grade 3	-	0.00	0.00
FracSegNet (Unweighted CE)	Grade 0	0.85	0.91	0.88
	Grade 1	0.10	0.08	0.09
	Grade 2	0.39	0.32	0.35
	Grade 3	0.00	0.00	0.00

Table 4.4: Comparison of classification metrics for FracNet, FracSegNet (Focal Loss), and FracSegNet (Unweighted CE) across different fracture grades. ”-” denotes undefined due to division by zero.

Chapter 5

Discussion

5.1 Vertebrae Localisation

From the vertebrae localisation results in Table 4.1, it is clear that the two models, ***LocNet-F*** and ***LocNet-H***, demonstrate contrasting performance characteristics across the cervical, thoracic, and lumbar regions of the spine.

Cervical Spine Localisation

In the cervical region, ***LocNet-H*** exhibits superior performance with a considerably lower mean localisation error of 1.19mm $\pm$ 0.5mm compared to 2.58mm $\pm$ 2.1mm for ***LocNet-F***. This outcome suggests that the heatmap-only approach is particularly effective in this region, where the vertebrae are densely packed and smaller than those in the thoracic and lumbar regions, which contrasts with the assertions made by Cui et al. [34].

We assert that the observed discrepancies in the cervical region's performance can be attributed to the dense arrangement of vertebrae in this area. As a result of the cervical vertebrae being closely packed, there is overlap between the localised high-intensity regions within the weighted vote map, M, generated by ***LocNet-F***. These high-intensity regions correspond to the areas where the vote maps for individual vertebrae converge, leading to high aggregated values that are not necessarily centered on the true vertebral bodies but are instead located at the intersections of adjacent vertebral vote map regions. Consequently, the fast peak search clustering algorithm tends to erroneously identify centers at the peripheries of these overlapping regions rather than at the true anatomical centers. This specific problem is visualised in Figure 5.1, where the image on the left is

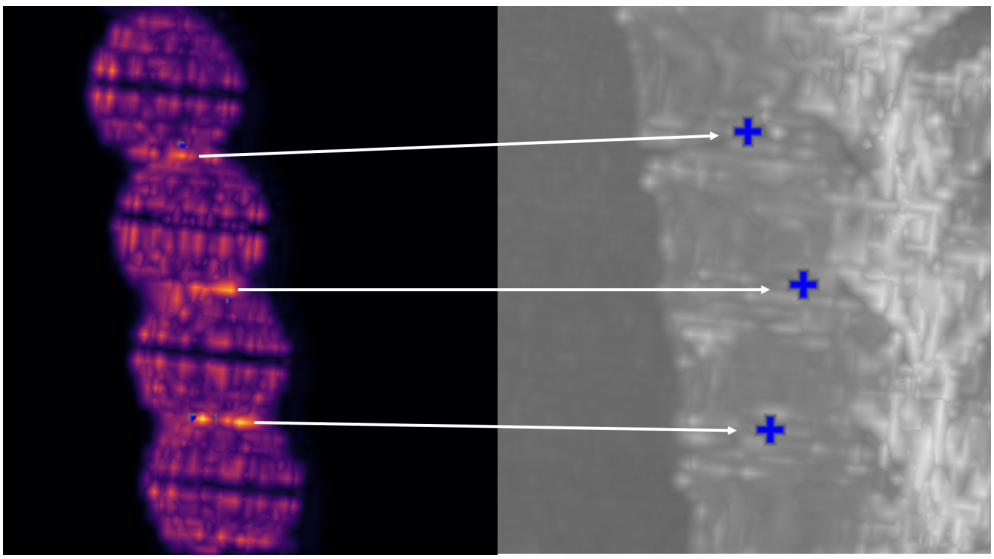

Figure 5.1: The left panel displays the weighted vote map, $\hat{M}$ for a typical case of cervical vertebrae, where brighter regions indicate higher aggregated values. The right panel shows the corresponding CT image with blue crosses marking the centers predicted by the fast peak search clustering algorithm. The overlap of vote maps between adjacent cervical vertebrae leads to inaccurate center predictions at the periphery of these overlapping regions, rather than at the true vertebrae centers.

the weighted vote map, M, and on the right is the corresponding CT image where the blue points denote the centers predicted by applying fast-search clustering on M. In the weighted vote map, the brighter appearing regions correspond to higher intensity values due to overlapping vertebral vote maps, which are consequently localised as centers by the clustering algorithm. In contrast, the heatmap-only approach, is unaffected by the issue of overlapping vote map regions. As a result, it yields more accurate localisation in the cervical spine. The reliance on direct intensity values from the heatmap allows for more accurate identification of the vertebral centers.

Consequently, the performance of ***LocNet-F*** in the cervical region could be improved by using Gaussian heatmaps with a smaller standard deviation, δ, during training. A smaller δ would produce smaller, more focused heatmaps, resulting in fewer foreground voxels being selected from the offset map for vote map generation, as outlined in Algorithm 1. This reduction in the number of selected voxels would necessarily reduce the size of each vertebra's vote map region, preventing spatial overlap between vote maps of adjacent vertebrae, ensuring that the votes in M are concentrated closer to the true anatomical centers rather than dispersed across regions between vertebrae. This adjustment would allow ***LocNet-F*** to better distinguish between closely spaced

vertebrae in the cervical region, improving the precision and consistency of vertebral center localisation.

Thoracic Spine Localisation

In the thoracic region, the localisation errors between ***LocNet-F*** and ***LocNet-H*** are similar, with ***LocNet-F*** exhibiting an error of 3.78mm, slightly higher than the 3.61mm achieved by ***LocNet-H***. The thoracic vertebrae, while not as densely packed as the cervical vertebrae, are more spatially separated, which mitigates the vote map overlap issue that affects the cervical region, as seen in a typical case involving thoracic vertebrae in Figure 5.2.

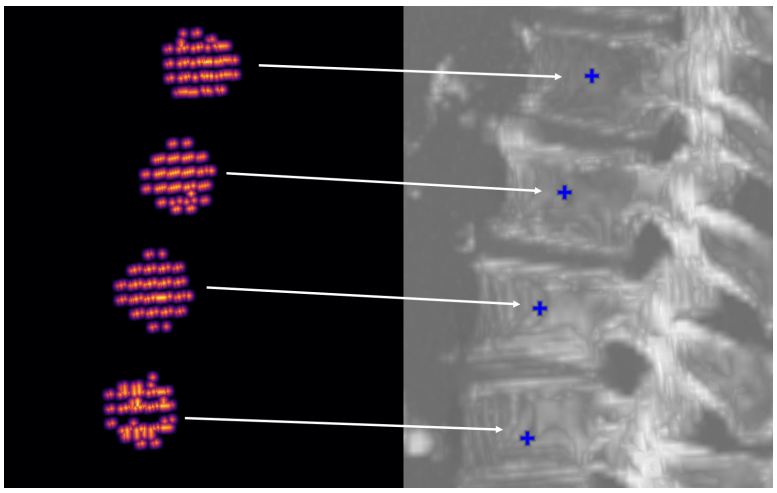

Figure 5.2: Comparison of the weighted vote map, $\hat{M}$ (left) and the corresponding CT image (right) for a typical case of thoracic vertebrae. In the vote map, brighter regions indicate higher aggregated values. The CT image shows blue crosses marking the centers predicted by the fast peak search clustering algorithm. Unlike in the case of cervical vertebrae, there is no overlap of vote maps, resulting in precise localisation.

This lack of overlap also enables ***LocNet-F*** to more accurately localise vertebral centers in the upper thoracic spine, where the vertebrae are still relatively densely packed. As shown in Figure 5.3, ***LocNet-H*** misses several upper thoracic vertebrae that are successfully localised by ***LocNet-F***. This discrepancy contributes to the lower standard deviation observed for the latter, as it consistently identifies more vertebrae, leading to a more stable error distribution. However, for vertebrae that are successfully localised by both models, ***LocNet-H*** achieves slightly more accurate localisation, as seen in the last vertebra in Figure 5.3.

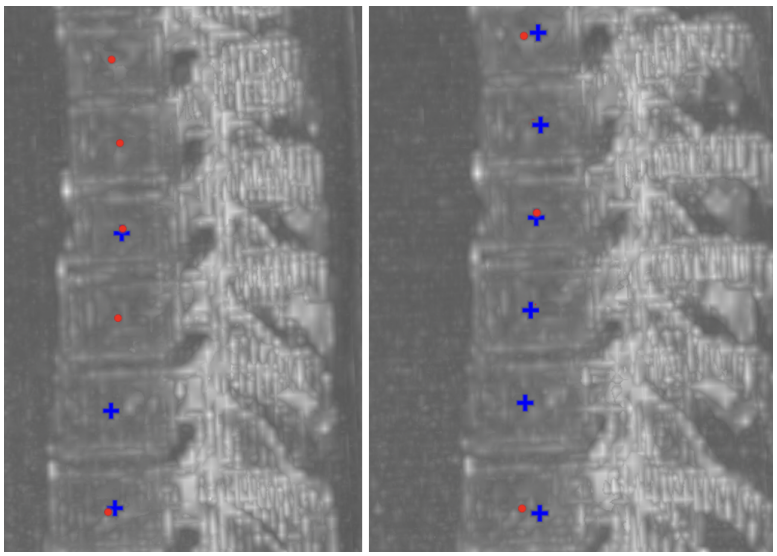

Figure 5.3: Comparison of vertebral center localisation in the upper thoracic spine using ***LocNet-H*** (left) and ***LocNet-F*** (right). ***LocNet-H*** misses several vertebrae in this region, whereas ***LocNet-F*** successfully localises them. Red dots indicate ground truth centers, and blue crosses indicate predicted centers.

Crucial to our concerns, the robustness of ***LocNet-F*** is particularly evident in cases involving fractures, where precise localisation is crucial. While ***LocNet-H*** shows a slightly lower mean localisation error, its higher standard deviation can be attributed to its inconsistency, especially in pathological cases where it may miss more vertebrae. In contrast, ***LocNet-F***'s use of the offset map and weighted voting mechanism results in more consistent localisation, even in challenging cases. As shown in Figure 5.4, ***LocNet-F*** successfully localises the center of a vertebra with a Grade 3 fracture, whereas ***LocNet-H*** completely misses it. This consistency demonstrates ***LocNet-F***'s robustness in handling difficult cases, leading to more reliable results across different thoracic scans, particularly in pathological scenarios.

Lumbar Spine Localisation

In the lumbar region, ***LocNet-F*** significantly outperforms ***LocNet-H***, with a localisation error of 3.09 mm $\pm$ 5.4 mm compared to 6.06 mm $\pm$ 14.6 mm for ***LocNet-H***. The lumbar vertebrae are larger and more widely spaced compared to those in the cervical and thoracic regions, resulting in no overlap between vote maps, which minimises the risk of erroneous vote accumulation. This spatial separation contributes to ***LocNet-F***'s lower standard deviation, as it consistently identifies vertebrae across the lumbar region.

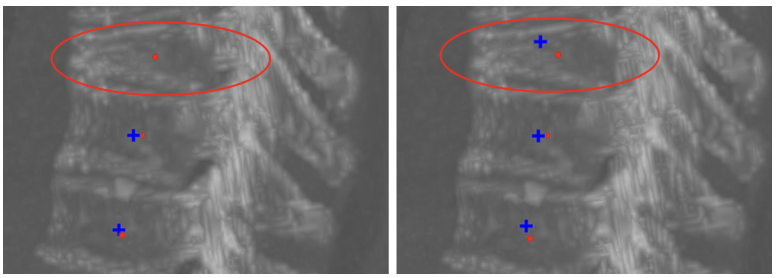

Figure 5.4: Comparison of thoracic vertebral center localisation in a pathological case involving a Grade 3 fracture. The left panel shows the results of ***LocNet-H***, where the vertebra with the fracture is missed entirely. The right panel displays the results of ***LocNet-F***, which successfully localises the center of the fractured vertebra.

The substantial standard deviation associated with ***LocNet-H*** in the lumbar region underscores the instability of the heatmap-only approach, particularly in managing the variability in vertebral size and orientation. This inconsistency can be attributed to ***LocNet-H***'s tendency to miss vertebrae, which leads to greater variability in the localisation error. Conversely, ***LocNet-F***'s integration of offset maps allows for more consistent localisation, as evidenced by the lower variance in error, demonstrating ***LocNet-F***'s superior adaptability to the morphology of the lumbar spine.

Overall Performance

When evaluating the performance across all spinal regions, ***LocNet-F*** consistently demonstrates superior accuracy, achieving a lower overall localisation error of 3.39mm $\pm$ 5.6mm, compared to **LocNet-H**'s 4.27mm $\pm$ 10.9mm. This indicates that the integration of offset maps and weighted vote mechanisms in ***LocNet-F*** offers significant advantages in achieving reliable and precise localisation. The lower standard deviation in ***LocNet-F*** further underscores its robustness, providing more consistent performance across varying anatomical complexities. Therefore, ***LocNet-F*** emerges as the more effective and dependable approach for vertebrae localisation, especially when considering its critical role as a precursor to accurate fracture detection.

5.2 Vertebrae Fracture Detection

With regard to vertebrae fracture detection, the breakdown of model performance, as depicted in Tables 4.2, 4.3 and 4.4 in Section 4.2, offers insight into the strengths and

limitations of all 3 models. Notably, the results across all three tables highlight the improved performance of ***FracSegNet*** with unweighted Cross-Entropy Loss over that of ***FracSegNet*** with Focal Loss. Consequently, we focus on the former model for the following discussion and going forward, refer to it simply as ***FracSegNet***.

The confusion matrix (Table 4.3) reveals a critical shortcoming of ***FracNet***, which successfully detects non-fractured vertebrae (grade 0) but fails entirely to identify higher-grade fractures (grades 1, 2, or 3), due to its inability to generalise during training. Although ***FracNet*** was trained using Focal Loss—specifically designed to mitigate such class imbalances—the network's inability to detect higher grades suggests that the information contained in the CT patches alone may be insufficient for accurate fracture classification, particularly for rarer fracture grades. Even after experimenting with different network depths, crop sizes, and both weighted and unweighted Cross-Entropy Loss, ***FracNet*** consistently showed poor performance, indicating that simply adjusting the architecture was not enough to overcome the model's limitations.

Furthermore, it could be argued that an alternative architecture, such as a 3D ResNet [69], which is well-regarded for its performance in image classification tasks, could have been employed to potentially achieve better classification results, as performed by Zhang et al. [38], albeit with a more balanced dataset. However, while the primary objective of this investigation was to perform classification of vertebral fractures, we also sought to explore whether integrating segmentation in a multi-task learning framework could enhance the classification of vertebral fractures. Given that U-Net is particularly adept at segmentation tasks, while also being suitable for classification tasks [70], it was chosen as the backbone architecture for both ***FracNet*** and ***FracSegNet*** to maintain consistency and to evaluate the potential benefits of shared representations with the same network architecture.

FracSegNet, while still imperfect, demonstrates some improvement over ***FracNet***. As shown in Table 4.2, ***FracSegNet*** is capable of identifying grade 2 fractures, and fractured vertebrae in general to a modest extent. The use of shared representations through multi-task learning appears to contribute to this enhancement, as it allows the model to leverage additional contextual information about the vertebra of interest from the segmentation task. This suggests that incorporating vertebral structure information alongside the CT data can provide more discriminative features for fracture detection.

The segmentation metrics indicate that for Grade 2 fractures, the model achieves a higher recall (0.93) but exhibits slightly lower precision (0.89) relative to Grade 1 and Grade 0 fractures. This suggests that while the model captures a greater proportion of

true positive voxels for Grade 2 fractures, it does so by accepting a higher rate of false positives in the segmentation. Despite this trade-off, the higher recall in segmentation contributes to the improved classification of Grade 2 fractures in the multitask ***FracSegNet***. The model's increased recall in identifying Grade 2 fracture voxels, despite lower precision, likely aids in the correct classification of these fractures by capturing more true positive voxels during segmentation. This observation aligns with findings by Oliveira et al. [71] who perform multi-task classification and segmentation of chronic venous disorders (CVD). In their work, classification accuracy was much more strongly correlated with segmentation recall than segmentation precision when leveraging a multi-task deep learning network that simultaneously performed both tasks. Conversely, for Grade 1 fractures, the model maintains a more balanced but less aggressive segmentation approach, with both precision and recall around 0.91. This balance likely leads to the underclassification of some Grade 1 fractures as Grade 0, as the model applies more stringent criteria during segmentation. In summary, the model's emphasis on recall over precision in Grade 2 fracture segmentation plays a crucial role in its better performance in classifying these fractures.

However, ***FracSegNet's*** continued struggle with Grade 3 fractures indicates that even multi-task learning is not fully sufficient to address the complexities of fracture detection in highly imbalanced datasets. In addition to the challenges associated with classifying vertebral fractures, the segmentation performance for Grade 3 fractures in the ***FracSegNet*** model is notably poorer compared to other grades, as evidenced by the lower segmentation metrics across the board. This disparity likely stems from the greater imbalance between foreground and background voxels in Grade 3 cases, where the affected vertebrae are significantly smaller and more fragmented compared to those with Grade 0, 1, or 2 fractures. This uneven distribution exacerbates the difficulty of accurate segmentation in Grade 3 fractures [72]. An example of this is shown in Figure 5.5, where the predicted segmentation mask generated by ***FracSegNet*** for a typical Grade 3 fracture demonstrates both over-segmentation and under-segmentation, with some background voxels being preferentially segmented, while some foreground voxels are not segmented. To address this issue, we propose that the adoption of Unified Focal Loss proposed by Yeung et al. [72], which is designed to manage such voxel-level imbalances, could improve the overall segmentation performance for vertebrae with Grade 3 fractures.

The Unified Focal Loss is a hierarchical loss framework that generalises Dice and cross entropy-based losses, specifically designed to manage class imbalance in medical

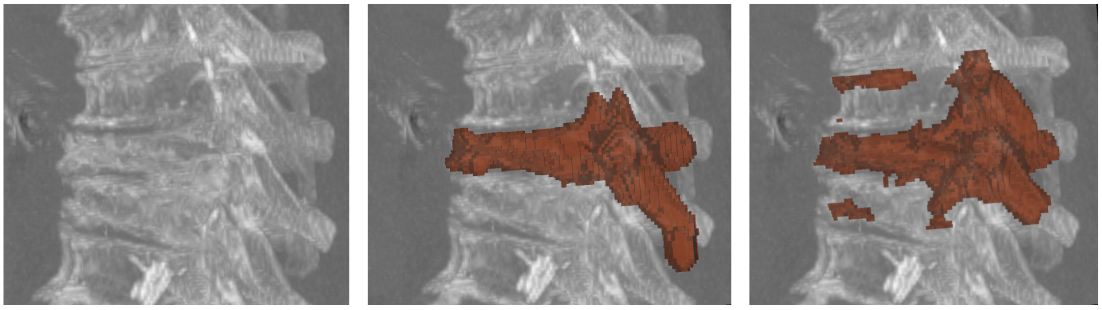

Figure 5.5: Visualisation of a grade 3 fractured vertebra (middle vertebra in each panel) in a CT crop (left panel). The middle panel shows the ground truth segmentation mask, while the right panel displays the predicted segmentation by ***FracSegNet***. The predicted segmentation mask exhibits over-segmentation and under-segmentation.

image segmentation tasks [72]. Unlike standard Dice loss, the Unified Focal Loss introduces a modulating factor that enhances learning from difficult, minority class samples while reducing the impact of easy, majority class examples. This is achieved by combining the strengths of the Focal Loss with a mechanism that balances recall and precision—specifically, controlling the trade-off between false negative voxels (under-segmentation) and false positive voxels (over-segmentation). By incorporating these elements into a unified framework, the Unified Focal Loss addresses the limitations of traditional loss functions, offering improved stability during training and better performance in handling class imbalances.

$$\mathcal{L}_{mF(p_t)} = \delta(1-p_t)^{1-\gamma} \cdot \mathcal{L}_{\text{BCE}}(p,y), \tag{5.1}$$

$$\mathcal{L}_{mFT} = \sum_{c=1}^{C} (1-\text{mTI})^{\gamma}, \tag{5.2}$$

$$\text{mTI} = \frac{\sum_{i=1}^{N} p_{0i}g_{0i}}{\sum_{i=1}^{N} p_{0i}g_{0i} + \delta\sum_{i=1}^{N} p_{0i}g_{1i} + (1-\delta)\sum_{i=1}^{N} p_{1i}g_{0i}}, \tag{5.3}$$

$$\Rightarrow \mathcal{L}_{UF} = \lambda\mathcal{L}_{mF} + (1-\lambda)\mathcal{L}_{mFT}, \tag{5.4}$$

where:

- $\mathcal{L}_{mF(p_t)}$ is a modified focal binary cross entropy loss, $\mathcal{L}_{\text{BCE}}$, that focuses on hard-to-classify examples by applying the modulation factor $(1-p_t)^{1-\gamma}$, where p_t represents the predicted probability for the true class [72].

- $\mathcal{L}_{mFT}$ is a modified focal Tversky Loss [73] that utilises the modified Tversky

Index (mTI) [74] to measure the overlap between the predicted and ground truth segmentation. The index is weighted by δ to prioritise false negatives or false positives.

- λ is a weighting factor that controls the balance between the two intermediate losses, $\mathcal{L}_{mF}$ and $\mathcal{L}_{mFT}$, allowing for fine-tuning based on the degree of class imbalance present in the dataset.

Applying this loss function to the segmentation task within ***FracSegNet*** could potentially improve the segmentation performance for grade 3 fractures, addressing the current limitation in the model's ability to accurately segment and classify these more severe fractures. Given that segmentation performance is intrinsically linked to the quality of the shared representations learned in the multi-task framework [75], enhancing segmentation through Unified Focal Loss could also positively impact the model's fracture classification capabilities, particularly for the rarer and clinically significant grade 3 fractures.

In conclusion, while multi-task learning represents a promising step forward compared to pure classification, particularly in its ability to leverage shared representations for fracture detection, overall performance remains limited, especially for the most severe fractures. More broadly, these challenges highlight the difficulty in developing robust methods for fracture detection in the face of the class imbalances inherent in medical imaging data, laying the groundwork for the proposed future work, which aims to address these limitations and further enhance the multi-task model's clinical utility.

Chapter 6

Future Work

Building on the findings and limitations of our methods in Chapters 4 and 5, we proceed to outline directions for future research. While multi-task learning has shown some potential in our experiments, as aforementioned, the inherent class imbalance in medical imaging datasets poses a continuous challenge in building robust automated tools for fracture detection. Despite the inclusion of Focal Loss for classification, current results indicate that additional strategies are necessary to fully realise the potential of a fracture detection model. To that end, we have proposed the use of Unified Focal Loss to improve segmentation performance in the class-imbalanced, multi-task learning setting, thereby enhancing the quality of shared representations and enabling robust generalisation on the aligned task of fracture detection [76]. Furthermore, investigating alternative loss functions that are sensitive to class imbalance in segmentation tasks should be an area of further research. For instance, Focal Tversky Loss [73] and Hybrid Loss [77] were specifically developed to address the challenges of class imbalance in medical image segmentation. Exploring these loss functions within the context of multi-task learning frameworks could potentially enhance segmentation accuracy, particularly for underrepresented classes, which in turn could lead to improvements in fracture classification performance.

By the same token, investigating different network architectures should also be prioritised. As aforementioned, a ResNet backbone could be implemented as the shared multi-task encoder to learn general representations. Such a network would subsequently have task-specific classification and segmentation decoders, where the latter incorporates a U-Net decoder that upsamples features to match the spatial dimensions of the input, as implemented by Graham et al. [76] for histology segmentation and classification. Conversely, the classification decoder would largely resemble the classification branch

in ***FracSegNet***, with the exception that it takes the output representation from the ResNet as input, instead of the output from a U-Net decoder.

Further expanding the multi-task framework, it would be pertinent to explore the addition of specialised branches for detecting vertebral endplate and posterior wall fractures, similar to the efforts of Zhang et al [38], though their work did not incorporate these tasks within a multi-task learning framework. Integrating these tasks into the same network could provide a more comprehensive assessment of vertebral health, aligning the model's output with the various ways in which vertebral fractures present. Detecting the presence of fractures in these specific regions could also contribute to more accurate overall fracture grade classification by offering additional context about the structural integrity of the vertebrae, thereby improving the model's ability to differentiate between fracture grades.

In conclusion, the proposed future work aims to improve fracture detection by refining both the technical approaches and the scope of the models. By addressing the current limitations and exploring new avenues in multi-task learning, segmentation, and network architecture, future research can move closer to developing more effective and reliable tools for clinical use.

Chapter 7

Conclusion

This thesis project introduced a two-stage pipeline for vertebral fracture detection, beginning with vertebrae localisation followed by a multi-task learning network to simultaneously perform vertebrae segmentation and fracture classification from 3D CT scans. The research aimed to address the critical issue of under-diagnosis of vertebral fractures, which can have severe consequences if left untreated. The integration of segmentation and classification tasks in a multi-task learning framework was hypothesised to improve diagnostic accuracy by leveraging shared representations of vertebral structures. While the multi-task learning approach (***FracSegNet***) did provide some benefits, particularly in improving the detection of grade 2 fractures compared to a purely classification-based model (***FracNet***), the overall performance remained insufficient for reliable clinical application. The results indicated that, although multi-task learning improved fracture detection accuracy in certain cases, the model struggled significantly with detecting and correctly classifying higher-grade fractures, especially grade 3, reflecting the inherent challenges posed by the class imbalance in the dataset.

Moreover, the quality of vertebral segmentation, particularly for grade 3 fractures, suggests that more targeted efforts are required to improve segmentation performance in the presence of severe fractures and imbalanced classes. As discussed in the proposed future work, incorporating loss functions such as Unified Focal Loss could potentially enhance the model's ability to accurately segment vertebrae affected by severe fractures. This, in turn, could improve the overall fracture detection accuracy, especially for underrepresented and clinically significant cases. Further research should prioritise these enhancements to address the limitations identified in the current model for the development of reliable, automated diagnostic tools that can be effectively used in clinical practice.

Bibliography

[1] Omar Hussain, Mayank Kaushal, Nitin Agarwal, Shekar Kurpad, and Saman Shabani. The role of magnetic resonance imaging and computed tomography in spinal cord injury. *Life (Basel)*, 13(8):1680, 2023.

[2] C.C. Wong and M.J. McGirt. Vertebral compression fractures: a review of current management and multimodal therapy. *Journal of Multidisciplinary Healthcare*, 6:205–214, 06 2013.

[3] D.L. Kendler, D.C. Bauer, K.S. Davison, L. Dian, D.A. Hanley, S.T. Harris, M.R. McClung, P.D. Miller, J.T. Schousboe, C.K. Yuen, and E.M. Lewiecki. Vertebral fractures: Clinical importance and management. *The American Journal of Medicine*, 129(2):221.e1–221.e10, February 2016.

[4] S. Paik, J. Park, J.Y. Hong, et al. Deep learning application of vertebral compression fracture detection using mask r-cnn. *Scientific Reports*, 14:16308, 2024.

[5] Jitender Manhas, Raj Kumar Gupta, and Partha Pratim Roy. A review on automated cancer detection in medical images using machine learning and deep learning based computational techniques: Challenges and opportunities. *Archives of Computational Methods in Engineering*, 29:2893–2933, 2022.

[6] Xiaoyan Liu, Li Qu, Zhenyu Xie, et al. Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation. *BioMed Engineering OnLine*, 23:52, 2024.

[7] Yashpal Kumar, Ashish Koul, Richa Singla, and Muhammad Farooq Ijaz. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, 14(7):8459–8486, 2023. Epub 2022 Jan 13.

[8] Charisma DeSai and Amit Agarwal. *Neuroanatomy, Spine*. StatPearls Publishing, Treasure Island (FL), 2023.

[9] Mayfield Clinic. Spine anatomy, anatomy of the human spine, Sep 2018.

[10] Nikolai Bogduk. Functional anatomy of the spine. *Handbook of Clinical Neurology*, 136:675–688, 2016.

[11] Lumen Learning OpenStax. Anatomy and physiology i.

[12] North Bristol NHS Trust. Spinal fractures, 2020.

[13] Judy Adams, Emma M. Clark, Gavin Clunie, Jill Griffin, Clare Groves, Kassim Javaid, Tim Jones, Sarah Leyland, Andrew Pearson, Nicola Peel, Opinder Sahota, Khalid Salem, Jo Sayer, Sonya Stephenson, and Virginia Wakefield. Clinical guidance for the effective identification of vertebral fractures, November 2017. National Osteoporosis Society.

[14] Matthew Varacallo, Travis J. Seaman, Jagmohan S. Jandu, and Peter Pizzutillo. *Osteopenia*. StatPearls Publishing, 2024.

[15] Yong Sang Park and Hyun Soo Kim. Prevention and treatment of multiple osteoporotic compression fracture. *Asian Spine Journal*, 8(3):382–390, Jun 2014.

[16] O. Yaman, M. Zileli, S. Şentürk, K. Paksoy, and S. Sharif. Kyphosis after thoracolumbar spine fractures: Wfns spine committee recommendations. *Neurospine*, 18(4):681–692, Dec 2021. Epub 2021 Dec 31.

[17] Christopher J. Donnally III, Christopher M. DiPompeo, and Matthew Varacallo. *Vertebral Compression Fractures*. StatPearls Publishing, Treasure Island (FL), updated 2023 aug 4 edition, 2024.

[18] J. Aso-Escario, C. Sebastián, A. Aso-Vizán, J. V. Martínez-Quiñones, F. Consolini, and R. Arregui. Delay in diagnosis of thoracolumbar fractures. *Orthopedic Reviews (Pavia)*, 11(2):7774, May 23 2019.

[19] Parth R. Patel and Olivia De Jesus. *CT Scan*. StatPearls Publishing, Treasure Island (FL), updated 2023 jan 2 edition, 2024.

[20] Michael C. Florkow, Kevin Willemsen, Vasco V. Mascarenhas, Edwin H.G. Oei, Maarten van Stralen, and Peter R. Seevinck. Magnetic resonance imaging versus

computed tomography for three-dimensional bone imaging of musculoskeletal pathologies: A review. *Journal of Magnetic Resonance Imaging*, 56(1):11–34, Jul 2022. Epub 2022 Jan 19.

[21] Emily Whitney and Anthony J. Alastra. *Vertebral Fracture*. StatPearls Publishing, Treasure Island (FL), updated 2023 apr 3 edition, 2024.

[22] NHS England. Statistical release - 21st july 2022, 2022. Accessed: 2024-07-23.

[23] Nahum Kiryati and Yehoshua Landau. Dataset growth in medical image analysis research. *Journal of Imaging*, 7(8):155, Aug 20 2021.

[24] E. Michael Lewiecki and Andrew J. Laster. Clinical applications of vertebral fracture assessment by dual-energy x-ray absorptiometry. *The Journal of Clinical Endocrinology Metabolism*, 91(11):4215–4222, November 2006.

[25] Harry K. Genant, Chih-Ying Wu, Chris van Kuijk, and Michael C. Nevitt. Vertebral fracture assessment using a semiquantitative technique. *Journal of Bone and Mineral Research*, 8(9):1137–1148, Sep 1993. PMID: 8237484.

[26] M. Grigoryan, A. Guermazi, F.W. Roemer, P.D. Delmas, and H.K. Genant. Recognizing and reporting osteoporotic vertebral fractures. *European Spine Journal*, 12 Suppl 2(Suppl 2):S104–S112, Oct 2003. Epub 2003 Sep 11. PMID: 13680316; PMCID: PMC3591834.

[27] Animesh Panda, Chandan J. Das, and Utpal Baruah. Imaging of vertebral fractures. *Indian Journal of Endocrinology and Metabolism*, 18(3):295–303, May 2014. Erratum in: Indian J Endocrinol Metab. 2014 Jul;18(4):581.

[28] Leon Lenchik, Lee F. Rogers, Pierre D. Delmas, and Harry K. Genant. Diagnosis of osteoporotic vertebral fractures: Importance of recognition and description by radiologists. *American Journal of Roentgenology*, 183(4):949–958, 2004. PMID: 15385286.

[29] M.G. Bendtsen and M.F. Hitz. Opportunistic identification of vertebral compression fractures on ct scans of the chest and abdomen, using an ai algorithm, in a real-life setting. *Calcified Tissue International*, 114:468–479, May 2024. Received 22 December 2023; Accepted 13 February 2024; Published 26 March 2024.

[30] Stefan Schmidt, Jörg Kappes, Martin Bergtholdt, Vladimir Pekar, Sebastian Dries, Daniel Bystrov, and Christoph Schnörr. Spine detection and labeling using a parts-based graphical model. *Lecture Notes in Computer Science*, page 122–133, 2007.

[31] Ben Glocker, J. Feulner, Antonio Criminisi, D. R. Haynor, and E. Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, page 590–598, 2012.

[32] Haofu Liao, Addisu Mesfin, and Jiebo Luo. Joint vertebrae identification and localization in spinal ct images by combining short- and long-range contextual information. *IEEE Transactions on Medical Imaging*, 37(5):1266–1275, 05 2018.

[33] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[34] Zhiming Cui, Changjian Li, Lei Yang, Chunfeng Lian, Feng Shi, Wenping Wang, Dijia Wu, and Dinggang Shen. Vertnet: Accurate vertebra localization and identification network from ct images. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, page 281–290, 09 2021.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*, page 234–241, 2015.

[37] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 124–133, 2020.

[38] J. Zhang, F. Liu, J. Xu, Q. Zhao, C. Huang, Y. Yu, and H. Yuan. Automated detection and classification of acute vertebral body fractures using a convolutional

neural network on computed tomography. *Frontiers in Endocrinology (Lausanne)*, 14:1132725, Mar 27 2023. PMID: 37051194; PMCID: PMC10083489.

[39] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[40] S.A. Nadeem, A.P. Comellas, E.A. Regan, E.A. Hoffman, and P.K. Saha. Chest ct-based automated vertebral fracture assessment using artificial intelligence and morphologic features. *Medical Physics*, 51(6):4201–4218, Jun 2024. Epub 2024 May 9. PMID: 38721977.

[41] F. Bragman, R. Tanno, Z. Eaton-Rosen, W. Li, D. Hawkes, S. Ourselin, D. Alexander, J. McClelland, and M.J. Cardoso. Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[42] Ryutaro Tanno, M.A. Arslan, O. Oktay, S. Mischkewitz, F. Al-Noor, J. Oppenheimer, R. Mandegaran, B. Kainz, and M.P. Heinrich. Autodvt: Joint real-time classification for vein compressibility analysis in deep vein thrombosis ultrasound diagnostics. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[43] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[44] Michael Crawshaw. Multi-task learning with deep neural networks: A survey, 2020.

[45] Derya Soydaner and Johan Wagemans. Multi-task convolutional neural network for image aesthetic assessment. *IEEE Access*, 12:4716–4729, 2024.

[46] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014.

[47] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[48] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 129(12):3174–3194, September 2021.

[49] Jingru Yi, Pengxiang Wu, Qiaoying Huang, Hui Qu, and Dimitris N. Metaxas. Vertebra-focused landmark detection for scoliosis assessment, 2020.

[50] Yu-Ching Yeh, Cheng-Han Weng, Yu-Jen Huang, Ming-Feng Tsai, Ming-Ting Wu, Sheng-Long Hung, and Hsiu-Po Hsu. Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs. *Scientific Reports*, 11(1):7618, 2021.

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization, 2017.

[52] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015.

[53] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2016.

[54] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[55] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[56] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision Workshops (ECCVW)*, 2016.

[57] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[58] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression, 2020.

[59] Anjany Sekuboyina, Mohamed Emad Husseini, Abolfazl Bayat, Markus Löffler, Helmut Liebl, Hengameh Li, Giles Tetteh, Jiří Kukačka, Christian Payer,

Darko Štern, Martin Urschler, Mingyuan Chen, Dazhou Cheng, Nils Lessmann, Yuhang Hu, Tianyu Wang, Dexing Yang, Dongnan Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Stefan Lehnert, Marcus Lirio, Nathaniel P Olaguer, Harry Ramm, Monika Sahu, Alexander Tack, Stefan Zachow, Tianming Jiang, Xin Ma, Craig Angerman, Xuanang Wang, Kristin Brown, Aida Kirszenberg, Émilien Puybareau, Dongfang Chen, Yu Bai, Brandon H Rapazzo, Timothy Yeah, Antonia Zhang, Shan Xu, Fang Hou, Zhijian He, Changhao Zeng, Zhu Xiangshang, Xu Liming, Timothy J Netherton, Richard P Mumme, Laurence E Court, Zhiwei Huang, Chao He, Lizhi Wang, Shaohua Ling, Linh Dan Huỳnh, Nicolas Boutry, Radim Jakubicek, Jan Chmelik, Shashank Mulay, Mohanasankar Sivaprakasam, Johannes C Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitsch, Markus Rempfler, Bjoern H Menze, and Jan S Kirschke. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical Image Analysis*, 73:102166, Oct 2021. Epub 2021 Jul 22.

[60] Image Biomarker Standardisation Initiative (IBSI). IBSI: Image Processing Report, 2024. Accessed: 2024-07-23.

[61] Amir Pasha Mahmoudzadeh and Nasser H. Kashou. Evaluation of interpolation effects on upsampling and accuracy of cost functions-based optimized automatic image registration. *International Journal of Biomedical Imaging*, 2013:19 pages, 2013.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[63] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[64] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2014.

[65] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[66] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[67] Jingyao Wu, Zhibin Zhao, Chuang Sun, Ruqiang Yan, and Xuefeng Chen. Learning from class-imbalanced data with a model-agnostic framework for machine intelligent diagnosis. *Reliability Engineering & System Safety*, 216:107934, 2021.

[68] O. U. Aydin, A. A. Taha, A. Hilbert, et al. On the usage of average hausdorff distance for segmentation performance assessment: hidden error when used for ranking. *European Radiology Experimental*, 5(4), 2021.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[70] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021.

[71] Bruno Oliveira, Hugo R. Torres, Paulo Morais, Francisco Veloso, Ana L. Baptista, José C. Fonseca, and José L. Vilaça. A multi-task convolutional neural network for classification and segmentation of chronic venous disorders. *Scientific Reports*, 13(1):761, January 2023.

[72] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95:102026, 2022.

[73] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687, 2019.

[74] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks, 2017.

[75] Weiwei Zhang, Guang Yang, Nan Zhang, Lei Xu, Xiaoqing Wang, Yanping Zhang, Heye Zhang, Javier Del Ser, and Victor Hugo C. de Albuquerque. Multi-task learning with multi-view weighted fusion attention for artery-specific calcification analysis. *Information Fusion*, 71:64–76, 2021.

[76] Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Shan E Ahmed Raza, Fayyaz Minhas, David Snead, and Nasir Rajpoot. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023.

[77] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy, 2021.

Appendix A

Localisation Network

A.1 Fast-search clustering for identifying peaks in M

Algorithm 2 Fast Search Clustering

Input: Weighted vote map M, Value threshold η, Distance threshold λ.

Output: List of density peaks P.

$P \leftarrow \emptyset$ ▷ Initialise list of peaks

for each voxel (z,y,x) in M **do**

  if $M[z,y,x] > \eta$ **then**

    neighborhood $\leftarrow M[z-1:z+2, y-1:y+2, x-1:x+2]$

    if $M[z,y,x] = \max(\text{neighborhood})$ **then**

      higher_density_voxels $\leftarrow \{v \mid M[v] > M[z,y,x]\}$

      if higher_density_voxels is empty **then**

        $P \leftarrow P \cup \{(z,y,x)\}$ ▷ Add isolated high value peak

        continue

      end if

      distances $\leftarrow \{\|v-(z,y,x)\| \mid v \in \text{higher\_density\_voxels}\}$

      if $\min(\text{distances}) > \lambda$ **then**

        $P \leftarrow P \cup \{(z,y,x)\}$ ▷ Add peak if sufficiently separated

      end if

    end if

  end if

end for

Return: P

A.2 VerSe2019 Dataset Statistics

	Training	**Validation**	**Test**
Number of CT images	80	40	40
Smallest CT image size	$114 \times 152 \times 76$	$103 \times 157 \times 76$	$129 \times 144 \times 68$
Largest CT image size	$915 \times 1189 \times 709$	$538 \times 702 \times 683$	$656 \times 733 \times 787$
Mean CT image size	$270.575 \times 339.2 \times 288.6$	$252.325 \times 350.725 \times 230.9$	$265.275 \times 340.05 \times 258.8$
Standard deviation of CT image sizes	$124.1189 \times 173.6119 \times 187.2345$	$104.7204 \times 169.5830 \times 141.7891$	$131.9945 \times 160.6758 \times 184.1612$

Table A.1: Statistics of CT image sizes for training, validation, and test datasets.