Vertebrae segmentation and labeling from CT images

Di Meng

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques et Informatique**

Arrêté ministériel : 15th November 2022

Présentée par

Di Meng

Thèse dirigée par **Edmond Boyer**
et codirigée par **Sergi Pujades**

préparée au sein **INRIA Grenoble Rhône-Alpes**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Vertebrae segmentation and labeling from CT images

Segmentation et étiquetage des vertèbres à partir d'images CT

Thèse soutenue publiquement le **15th November 2022**,
devant le jury composé de :

Prof. Dr. Edmond Boyer
INRIA Grenoble Rhône-Alpes, Directeur de thèse
Prof. Dr. Sergi Pujades
INRIA Grenoble Rhône-Alpes, Co-Directeur de thèse
Prof. Dr. Diana Mateus
Ecole Centrale Nantes, Rapporteur
Prof. Dr. Su Ruan
University of Rouen Normandy, Rapporteur
Prof. Dr. Jocelyne Troccaz
Université Grenoble Alpes, CNRS, Président
Dr. Benjamin Aubert
EOS image, Inc., Examinateur

Abstract

Human spine is a significant body structure that allows human to move freely and bend with flexibility. It also protects the spinal cord, a column of nerves connecting the brain with the rest of the body. However, the injury, health, lifestyle or other factors may affect the spine condition and cause spine diseases and disorders. An early diagnosis and treatment is demanded for the prevention and rehabilitation of the spine diseases. Computerized spine analysis has become essential in spine diagnosis and computer-assisted spine surgery interventions. Obtaining a reliable spine model with accurate vertebrae structures is a prerequisite. This thesis concentrates on this task, namely vertebrae localization, segmentation and identification from CT images. Existing works have developed conventional or learning-based approaches towards this task. There still remains some challenges, such as the detection of transitional vertebrae and the pathological cases where they are poorly present in the training dataset. Our work employs the combination of statistical shape priors and the deep networks to tackle the problem. In this thesis, we present the following contributions: We first learn a statistical surface model of the full spine from partial or incomplete observations using Probabilistic Principle Component Analysis (PPCA). The obtained spine model faithfully captures the shape of the vertebrae and is able to predict the unseen vertebrae given a few observations. We further contribute a combined local and global strategy for vertebrae identification. The individual vertebra label is predicted by an hierarchical-architecture network taking its morphology as input. A graphical model is designed to enforce the consistency over the individual predictions. The graph also explicitly models the transitional vertebrae which well detects the presence of T13, L6 and the absence of T12. In the end, we propose an anatomic consistency cycle to obtain a unified and coherent result of vertebrae locations, segmentation and identifications. It leverages the statistics of the shape priors and the deep networks, coping with the rarely abnormal cases which can be neglected by data-driven approaches. All proposed methods are validated on datasets both qualitatively and quantitatively. The derived model and code are made available to the community.

Résumé

La colonne vertébrale humaine est une structure corporelle importante qui permet à l'homme de bouger librement et de se plier avec souplesse. Il protège également la moelle épinière, une colonne de nerfs reliant le cerveau au reste du corps. Cependant, les blessures, la santé, le mode de vie ou d'autres facteurs peuvent affecter l'état de la colonne vertébrale et provoquer des maladies et des troubles de la colonne vertébrale. Un diagnostic et un traitement précoces sont nécessaires pour la prévention et la réhabilitation des maladies de la colonne vertébrale. L'analyse informatisée du rachis est devenue essentielle dans le diagnostic du rachis et l'intervention assistée par ordinateur en chirurgie du rachis. L'obtention d'un modèle de colonne vertébrale fiable avec des structures de vertèbres précises est une condition préalable. Cette thèse se concentre sur cette tâche, à savoir la localisation, la segmentation et l'identification des vertèbres à partir d'images CT. Les travaux existants ont développé des approches conventionnelles ou basées sur l'apprentissage pour cette tâche. Il reste encore quelques défis, tels que la détection des vertèbres de transition et les cas pathologiques où elles sont mal présentes dans l'ensemble de données d'entraînement. Notre travail utilise la combinaison des lois a priori statistiques et des réseaux de neurones profonds pour résoudre le problème. Dans cette thèse, nous présentons les contributions suivantes : Nous apprenons d'abord un modèle de surface statistique de la colonne vertébrale complète à partir des observations partielles ou incomplètes à l'aide de l'analyse probabiliste des composants principaux (PPCA). Le modèle de colonne vertébrale obtenu capture fidèlement la forme des vertèbres et est capable de prédire les vertèbres invisibles à partir de quelques observations. Nous contribuons en outre à une stratégie combinée locale et globale pour l'identification des vertèbres. L'étiquette de vertèbre individuelle est prédite par un réseau d'architecture hiérarchique prenant sa morphologie en entrée. Un modèle graphique est conçu pour renforcer la cohérence des prédictions individuelles. Le graphique modélise également, de manière explicite, les vertèbres de transition qui détectent bien la présence de T13, L6 et l'absence de T12. Au final, nous proposons un cycle de cohérence anatomique pour obtenir un résultat unifié et cohérent des localisations, segmentations et identifications des vertèbres. Il exploite les statistiques des a priori de forme et des réseaux de neurones profonds, faisant face aux cas rarement anormaux qui peuvent être négligés par les approches basées sur les données. Toutes les méthodes proposées sont validées sur des ensembles de données à la fois qualitativement et quantitativement. Le modèle et le code dérivés sont mis à la disposition de la communauté scientifique.

ACKNOWLEDGMENT

First and foremost I would like to thank Dr. Edmond Boyer and Dr. Julien Pansiot for having me in the team so that I had the opportunity to start my PhD journey.

My sincere gratitude goes to my supervisor Dr. Edmond Boyer. His open, curious and tolerant attitude towards research lets me perceive a deeper side of doing the research. I got inspired and also reassured each time after having a discussion with him. Among all his deliveries, I like most the saying of "A negative result is a result".

I also would like to express my gratitude to my co-supervisor Dr. Sergi Pujades. He provided a lot of advises and devoted plenty of time for my work. I learned how to research and present the research work from him. He also introduced me to interesting workshops which was a great experience to me.

I am also grateful to work with other collaborators of the project, Dr. Lucie Thiébaut from eCential Robotics, the surgeon Dr. Aurélien Courvoisier from CHU and other people from the project. They expanded my horizon of the industrial and clinical world. I obtained a novel experience collaborating with them.

I would like to give my thanks to INRIA. The institute provided me sufficient facilities for the research, especially during the COVID19 confinement time. I truly enjoyed this working environment. Also, I will miss the time I spent with my colleagues (the former and present team members) in the cafeteria, with the mountains around.

In the end, I would like to express my gratitude to my family. 感谢爸爸妈妈常年的理解和支持，让我毫无顾虑的专心学业。感谢舅舅舅妈和哥哥姐姐们在海外对我的照顾和关心。感谢身边朋友的陪伴。你们的存在是支持我完成此博士论文的最大动力。

CONTENTS

CHAPTER 1

INTRODUCTION

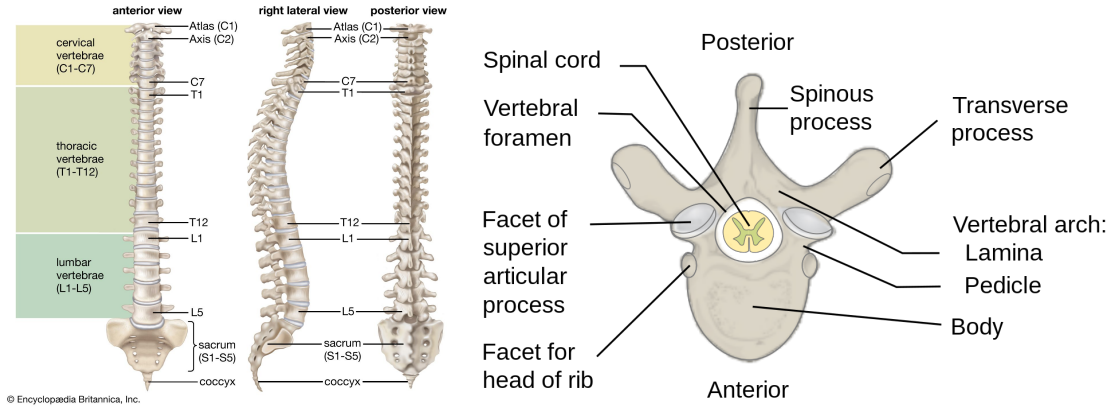
1.1 GENERAL CONTEXT

The human spine, also known as the vertebral column, normally is consisted of 33 vertebrae [107]. The upper 24 pre-sacral vertebrae are articulating and separated from each other by inter-vertebral discs. The lower 9 vertebrae are fused in adults, 5 in the sacrum and 4 in the coccyx. The articulating vertebrae are named according to their region and position of the spine. There are 7 cervical vertebrae, 12 thoracic vertebrae and 5 lumbar vertebrae. The number of the thoracic and lumbar vertebrae may vary in some population. An illustration of the spine and the individual vertebra is shown in Figure 1.1.

A number of conditions and injuries can affect the spine, which can damage the vertebrae and the inner spinal cord and nerves that they protect, cause pain, and limit mobility. For instance, a sedentary lifestyle or improper lifting techniques can cause neck pain or low back pain. Poor posture leads to a potential scoliosis especially in teenage group. Vertebral fractures can come from excessive and repetitive strenuous activities or other risk factors. Other spine disorders such as degenerative discs, osteoporosis, kyphosis and spinal deformities are affecting human's health and needs care and attention.

Medical imaging techniques are widely used for spine diagnosis and computer-assisted surgical interventions. Magnetic resonance (MR) imaging is recognized as the imaging technique of choice for accessing the inter-vertebral disc degeneration and abnormalities due to its excellent soft tissue contrast [37]. While computed tomography (CT) is broadly used in the examination of vertebral fractures and spinal curvatures as it presents high contrast for bony structures and requires less acquisition time (3-7 minutes). In this manuscript, we work with CT images and focus on the spine and vertebrae.

To have a basic concept of the CT, a CT scan is a medical imaging technique used in radiology to obtain detailed internal images of the body noninvasively for diagnosis purposes. CT scanners use a rotating X-ray tube and a row of detectors placed in the gantry to measure X-ray attenuations by different tissues inside the body. The multiple X-ray measurements taken from different angles are then reconstructed to produce tomographic



(a) The spine [2]

(b) The vertebra [3]

Figure 1.1: Anatomy of the spine and the vertebra.

images (namely slices) of a body. The slice thickness refers to the axial resolution of the scan. A series of slices can form a 3D CT volume through some registration algorithm that provides structural information of the scanned object. The radiodensity of the CT image is described by Hounsfield units (HU), which is a dimensionless unit to express CT numbers in a standardized and convenient form. Hounsfield units are obtained from a linear transformation of the measured attenuation coefficients. The transformation is based on the arbitrary-assigned densities of air and pure water that, radiodensity of distilled water at standard temperature and pressure (STP) is equal to 0 HU and radiodensity of air at STP is -1000 HU. Knowing the average linear attenuation coefficient of a body tissue or material, its corresponding HU value is therefore given by:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{1.1}$$

The exact HU of a body tissue can vary from one CT acquisition to another due to CT acquisition and reconstruction parameters. It is always in a constant range. For example, the fat HU is from -120 to -90. HU of soft tissue is from $+100$ to $+300$. Cancellous bone is in range from $+300$ to $+400$ while cortical bone HU is from 500 to 1900. The HU of metals is over 3000. The bone, seen as a high density tissue, absorbs the radiation to a greater degree, thus having a higher HU. On the contrast, the low density tissues, such as lungs, absorb the radiation to a lesser degree, thus having a lower HU.

A CT scan encodes the patient's anatomical position. Anatomical position, or standard anatomical position, refers to the specific body orientation used when describing an individual's anatomy. The standard anatomical position of the human body consists of the body standing upright and facing forward with the legs parallel to one another. The upper limbs, or arms, hang at either side and the palms face forward, as illustrated in Figure 1.2. The standard anatomical position provides a clear and consistent way of describing human anatomy and physiology. When accessing an individual's anatomy, many anatomical terms are used to described the relative positioning of various appendages in relation to the standardized position. Such terms include posterior or inferior, which means towards

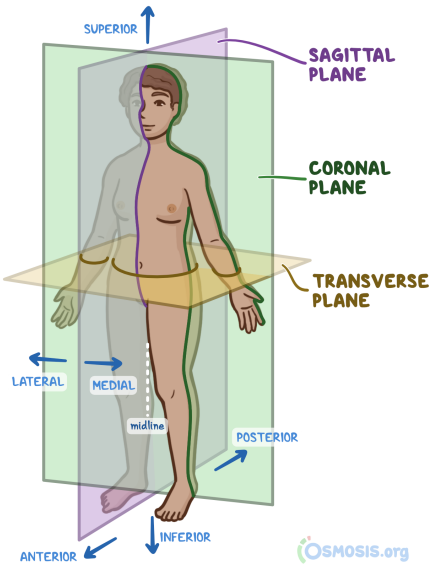


Figure 1.2: Standard anatomical position [9].

the back of the body, or towards the bottom of the body. Another imaginary reference is the three planes of the body, there are sagittal plane, coronal plane and the transverse (axial) plane, as shown in Figure 1.2. The sagittal plane is a vertical plane that travels down the middle of the human body and divides the body into right and left portions. The coronal plane runs vertically through the side of the body, dividing it into the front of the body and the back of the body. The transverse (axial) plane runs horizontally, separating the top half of the body from the bottom half. The standard anatomical position is encoded in the CT scan. It creates clear points of reference and helps to regularize the image orientation when processing multiple CT scans.

Having a spine model from CT scan is essential in many medical and clinical applications, such as detection of vertebral fractures [122], assessment of spinal deformities [39], study of anatomical structure [12], treatment planning [30] and computer-aided surgery [111, 63]. Computerized image segmentation has played an increasingly important role in the past decades. However, segmenting the spine and vertebrae from CT images and identifying them remain many challenges: i) highly varying field of views across the data. The view of a CT is usually narrowed down to a specific body part for a dedicated diagnosis; ii) large scan sizes. A CT volume can consist of up to 400 million voxels; iII) according to the nature of the vertebrae, neighboring vertebrae have a similar shape. It is difficult to tell apart one vertebra from the neighboring ones; iv) the number of vertebrae can change among the population, i.e. while most human have 12 thoracic vertebrae, some have 11 or 13. Similarly, some humans have 6 lumbar vertebrae instead of 5. These transitional vertebrae are found and reported to affect between 15% and 35% of the general population [21, 113, 65]; v) pathological spines, e.g. suffering from bone calcification or broken vertebrae, present pathological shapes; vi) given a random field of view CT, counting vertebrae is suboptimal [70] when no distinctive vertebra is available in the observation; vii)

the complex posterior elements of the vertebra makes the fine segmentation difficult; viii) the quality of the CT image is corrupted and noisy when there is metal implant present.

In this manuscript, we mainly focus on one major problem - vertebrae segmentation, localization and identification from CT images, which includes but not restricted to the sub-problems of learning a statistical full spine model from partial observations, spine segmentation, vertebrae localization, individual vertebra classification, transitional vertebrae detection and individual vertebra segmentation.

1.2 SCIENTIFIC CONTEXT AND MOTIVATION

This work was established at INRIA (French national research institute of computer science and automation), the MORPHEO team, directed by Dr. Edmond Boyer. The main research of the team focuses on the inner and outer human structure. In terms of inner human structure, organs or bones from medical volumetric images are analyzed. Regrading the outer human structure, surface model of the human is built and reconstructed statically and dynamically.

This work was established as part of the Spine PDCA project, jointly supported by the Auvergne Rhône Alpes region, the Fonds Unique Interministériel and the Grenoble Alpes metropole [grant number 1701595901-N00013834]. The project was conducted with four primary partners: eCential Robotics - a surgical device company, EOS imaging - a medical device company, CHU Grenoble Alpes - a central hospital, and INRIA - a research institute. The project targets to obtain a spine model from cone beam computed tomography (CBCT) in the context of scoliotic spine surgery and periodic spine curvature monitor for teenagers.

Cone beam computed tomography (CBCT) is a medical imaging technique similar to the traditional CT, consisting of X-ray computed tomography where the X-rays are divergent, forming a cone. CBCT has become increasingly important in treatment planning and diagnosis for its shorter acquisition time (2-3 minutes) and lower radiation exposure compared with CT. It is already widely used for oral and maxillofacial radiology. In this project, CBCT is employed for guiding the spine surgery and keep track of the scoliotic spine curvature. For this clinical usage, an automatic algorithm of obtaining the spine model from CBCT is in need. However, CBCT has its drawbacks, as the radiation dose is largely reduced, the image quality presents poor and noisy. The bone contrast is less clear compared with that in CT image. Manual segmenting the spine from CBCT is a difficult task even for a specialist. Therefore, the available annotated CBCT data is rare. Supervised learning segmentation strategy is not possible to be conducted. While there is relatively more annotated spine CT data in the community, the idea of transferring the learned algorithm of segmenting the spine from CT data to CBCT data is then proposed. Though the CBCT does not accurately respect the Hounsfield scale and the image intensity cannot be identical as CT, CBCT and CT share similarities in both image appearance and the scanned object features. Nowadays, cross-model domain adaptation is being investigated and developed [23, 75]. The objective is to apply the model learned with a

labeled dataset to an unlabeled dataset by minimizing the domain shift. To this purpose, the prerequisite of segmenting the spine from CBCT is to develop a robust method of segmenting the spine from CT images. This motivates us to work on building a spine model using CT data and develop a framework of vertebrae segmentation, localization and identification in CT images.

1.3 CONTRIBUTIONS

The contributions of this thesis are as follows:

- We proposed to learn a statistical surface model of the full spine from partial and incomplete views of the spine. The Probabilistic Principle Component Analysis (PPCA) is used to deal with the partial observations. In the method, the learned model never observed a complete full spine. Quantitative evaluation demonstrates that the obtained model faithfully capture the shape of the population in a low dimensional space and generalizes to left out data. It can be applied to predict the shape of unseen vertebrae with a mean distance error under 3mm. The model is made publicly available to the community for other applications.
- We proposed to use individual vertebra morphology for the vertebra classification and an upgraded version of including the contextual information. Quantitative evaluation shows that the shape of the individual vertebra can well represent the implicit vertebra features and is adequate to be used for the vertebrae distinction. The performance of solely using the vertebrae morphology is comparable to or higher than using the image appearance of the vertebrae.
- We proposed a hierarchical individual vertebra classification framework, where the anatomic group of the vertebra is firstly recognized and the individual vertebra label is then classified. The experiments show that the hierarchical classification architecture largely reduces the intra-class variance and achieves higher classification accuracy than directly classifying the vertebra into the individual class.
- We proposed a combined local and global approach for vertebrae identification. As the individual vertebra prediction from a local patch may have errors, we further proposed a graphical model for global reasoning. The graph enforces the anatomic consistency and select an optimal configuration given the predicted individual probabilities. Additionally, the graph explicitly models the transitional vertebrae that the presence of L6, T13 and the absence of T12 are well detected.
- We proposed to leverage the shape statistics with the deep networks for the unified task of vertebrae segmentation, localization and identification. For the rare cases such as transitional vertebrae, fractured or metal-inserted vertebrae or other pathological cases, they are hard to be learned from the limited data. We proposed an anatomic consistent cycle which combines both the deep learning strategy and the statistical priors. The vertebrae are iteratively localized, segmented and identified

with the anatomic consistency enforced. The method is benchmarked and achieved the state of the art performance. The code and the models are made publicly available for research purposes.

1.4 DATASETS

There are publicly available CT datasets with spine and vertebrae annotated. We choose to use the VerSe dataset to establish and validate our method, as it is currently the largest spine CT dataset. Further we use more datasets to test the robustness of the method. In the following, we introduce each of them.

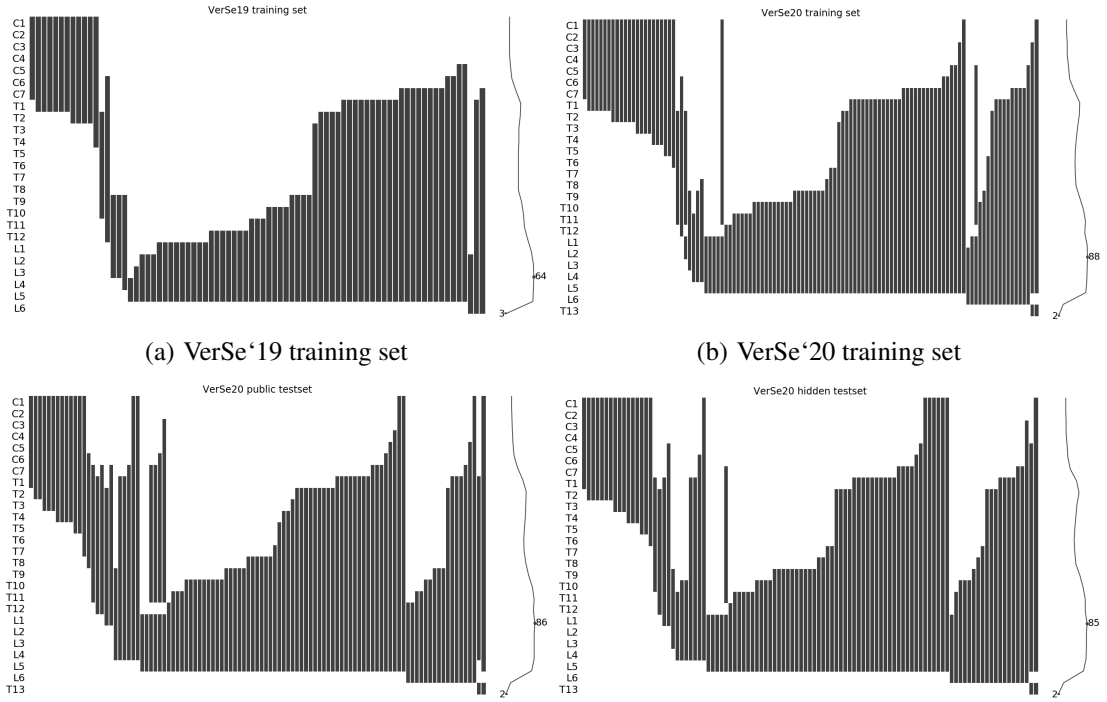
VerSe. [102] It is a publicly available dataset as part of the *Large Scale Vertebrae Segmentation* challenge, organized in conjunction with MICCAI 2019 and 2020. The challenge calls for fully-automated and interactive algorithms for tasks of vertebrae labelling and vertebrae segmentation. The entire VerSe dataset consists of 374 CT scans, which 160 scans were released as part of VerSe'19 and 355 scans for VerSe'20. They are publicly available after anonymisation (including defacing).

The data was collected from 355 subjects with a mean age of $\sim 59(\pm 17)$ years. The data is multi-site was acquired using multiple CT scanners, including the four major manufacturers (GE, Siemens, Phillips and Toshiba). Care was taken to compose the data to resemble a typical clinical distribution in terms of fields-of-view (FoV), scan settings, and findings. For example: it consists of a variety of FoVs (including cervical, thoracolumbar and cervico- thoraco-lumbar scans), a mix of sagittal and isotropic reformations, and cases with vertebral fractures, metallic implants, and foreign materials. Figure 1.3 shows the dataset distribution.

The dataset consists of two types of annotations: 1) 3D coordinate locations of the vertebral centroids and 2) voxel-level labels as segmentation masks. Twenty six vertebrae (C1 to L5, and the transitional T13 and L6) were considered for annotation with labels from 1 to 24, along with labels 25 and 28 for L6 and T13, respectively. The annotations are generated using a human-computer hybrid approach, in which the initial centroids and segmentation masks are computed using an automated algorithm and then manually and iteratively corrected and refined.

VerSe	Patients	Scans	Scan split	Vertebrae(Cer/Tho/Lum)
2019	141	160	80/40/40	1725(220/884/621)
2020	300	319	113/103/103	4141(581/2255/1305)
Total	355	374	141/120/113	4505(611/2387/1507)

Table 1.1: Data-split and additional details concerning the two iterations of VerSe. Scan split indicates the split of the data into train/Public test/Hidden-test phases. Cer, Tho, and Lum refers to the number of vertebrae form the cervical, thoracic, and lumbar regions, respectively. Note that VerSe'20 consists some cases from VerSe'19, resulting in the total patients not being an ad hoc sum of the two iterations.



(a) VerSe‘19 training set (b) VerSe‘20 training set

(c) VerSe‘20 public testset (d) VerSe‘20 hidden testset

Figure 1.3: Overview of the VerSe‘19 and VerSe‘20 datasets. Each column represents a scan of a patient, and each row represents the presence of one of the 26 vertebrae, from C1 to L6. On the right the aggregated number of vertebrae is shown.

The dataset is split into three phases, both for the challenge setting and benchmarking. Three phases indicate the training set, public test set and hidden test set. The CT scans and their annotations were both accessible for the training set. Only CT scans were accessible for public test set and participants had no access to neither CT scans nor their annotations for hidden test set. Table 1.1 shows the splits and the numbers of cervical, thoracic and lumbar vertebrae respectively.

Lumbar vertebra segmentation CT image database [50] consists of 10 CT lumbar spine images of 10 healthy subjects, each containing 5 lumbar vertebrae (i.e. levels from L1 to L5). For each vertebra, a reference manual segmentation is provided in the form of a binary mask.

xVertSeg database [49] is the dataset for Segmentation and Classification of Fractured Vertebrae. It is released as part of the xVertSeg cahllenge in MICCAI 2016. The dataset contains 25 CT lumbar spine images including few non-fractured vertebrae and vertebrae with fractures of different morphological grades and cases. Among the 25 cases, 15 have a reference vertebra segmentation mask of lumbar vertebrae (L1-L5) and fracture classification, while for the other 10 no annotations are publicly provided.

CSI2014 segmentation dataset is released as part of the spine segmentation challenge in MICCAI 2014. It consists of 15 CT scans of healthy subjects. The thoracic and lumbar

vertebrae are observed (17 vertebrae in total) and their associated ground truth segmentation masks are provided. The dataset is officially split into training set of 10 CT scans and test set of 5 scans by the challenge organizer.

EOS Imaging test set is from our collaborated medical device company. It consists of 3 CT scans covering the full spine from C1 to L5. The ground truth segmentation is generated by the professional technicians. There are two subjects with contrast agents injected in their spinal cords. They are 74 and 60 years old respectively. The third subject is a 28 year-old young adult with moderate scoliosis.

CHU test set is from our collaborated hospital. It is a large dataset consisting of 675 CT subjects in various health conditions. The ground truth segmentation is generated by the doctors. The dataset is with rich field of views while all the scans include the abdomino-pelvic area.

1.5 MANUSCRIPT STRUCTURE

This manuscript is organized as follows. Chapter 2 presents a statistical full spine model that learned from partial observations using a Probabilistic Principle Component Analysis (PPCA). In Chapter 3, we present the spine binary segmentation from CT images and the initial exploration of multi-vertebrae segmentation, which inspires the next chapters of individual vertebra identification and segmentation. In Chapter 4, we demonstrate a combined local and global approach for vertebrae identification using the morphology and a graphical model. In Chapter 5, we focus on the full task of vertebrae segmentation, localization and identification. We show an anatomically coherent result can be obtained by leveraging the anatomic prior knowledge and the deep networks. We further assess the generalization capability and the limitation of the proposed framework on other publicly available datasets and in-house datasets in Chapter 6. In the end, we conclude the work of this manuscript and discuss the potential future work in Chapter 7.

CHAPTER 2

SPINE MODEL

2.1 INTRODUCTION

A reliable spine model with accurate vertebrae structures is essential in numerous medical applications such as image segmentation, orthopedics, anesthesiology and pathology quantification. A 3D spine model is necessary to diagnose and evaluate the severity of spinal deformities. For example, clinical indexes such as the orientation of the plane of maximal curvature or the spine torsion rely on the availability of 3D spine models [108]. Furthermore, these 3D models are also used to plan and evaluate outcomes of orthopedic treatments [93, 116]. A full spine model can also help segment 3D volumetric images such as computed tomography (CT) or magnetic resonance imaging (MRI), as well as radiographs like X-ray images, for instance by injecting prior knowledge on the global shape of the spine as a regularizer. Statistical models have shown their interest in the generation of synthetic data for different modalities. A full spine model can also be used as a reference structure in clinical practice supporting the localization of other organs, as well as be used in the diagnosis of spine scoliosis.

In order to create a model of the full spine using the classical methodology, the full spine of several subjects needs to be observed. However, observations of the full spine (e.g. CT images) are rare. Practitioners usually narrow down the field of view of the spine to one specific level for a detailed diagnosis [90, 36] as well as to reduce the dose given to the patient. Thus most work has been done to model the shape of the vertebrae which are consistently observed in the scans.

To build a full spine model, we propose to learn a statistical surface model of the full spine (7 cervical, 12 thoracic and 5 lumbar vertebrae) from partial and incomplete views of the spine. In order to deal with the partial observations, we use Probabilistic Principle Component Analysis (PPCA) to learn a surface shape model of the full spine. Quantitative evaluation demonstrates that the obtained model faithfully captures the shape of the population in a low dimensional space and generalizes to left out data. Furthermore, we show that the model faithfully captures the global correlations among the vertebrae

shape. Given a partial observation of the spine, i.e. a few vertebrae, the model can predict the shape of unseen vertebrae with a mean error under 3mm. The full-spine statistical model is trained on the VerSe'19 public dataset and is made publicly available to the community for non-commercial purpose[1].

2.2 Related work

The statistical shape model is one of the most employed statistical atlas [45]. Creating statistical models of shapes using surface measurements has attracted a lot of attention since the early work of Blanz and Vetter [13]. Lorenz et al [80] use Principle Component Analysis (PCA) to obtain a statistical shape model. They then generate a template shape and fit the template shape to the rest objects by a rough landmark based shape morphing and mesh relaxation. Rajamini et al. [96] share a similar idea of constructing a statistical model using PCA and fit the deformable shape model to the sparse 3D data by computing a Mahalanobis distance weighted least square fit. While early methods mostly rely on PCA, non-linear methods have been also proposed to overcome the fact that unseen shapes cannot be expressed by linear combinations of the training samples [106, 82].

2.2.1 Learning statistical shape models from incomplete data

In the medical domain, either the data is scanned partially only to capture the essential part of the structure while minimizing the radiation exposure, or the data is captured incomplete because of a pathology. Many work proposed strategies to learn from the incomplete data and partial observations.

Lüthi et al. [81] build a statistical shape model (SSM) of skull from a dataset with corrupted shapes. They divide the training surfaces into patches and each patch is assigned with a probability being an outlier. PCOut [38] is performed to detect the outliers. Probabilistic PCA (PPCA) is applied to the remaining set to build the SSM. However, splitting surfaces into patches has a disadvantage that a single corrupted landmark causes a complete patch to be excluded. losing the information of correct correspondences. Gutierrez et al. [43] avoid doing the split, they propose to use Robust PCA (RPCA) [20] as a tool for outlier detection and correction, by modeling the corrupted data matrix as the addition of a low-rank matrix containing the corrected data and a sparse matrix signaling outliers and missing data. Zhang et al. [125] also build a shape composition model from sparse and erroneous data using a sparse learning method, where they assume the training shapes can be represented as a linear combination. However the distribution of the shape variance of biological structures is often non-linear and cannot be assumed as a Gaussian distribution. For example, the mean shape a single vertebra is not able to represent any existing vertebra [61]. A generic SSM is introduced by Ma et al. [82] to model non-linear distribution using Robust Kernel PCA (RKPCA). They construct a low-rank nonlinear subspace where outliers are discarded.

[1] `https://gitlab.inria.fr/spine/spine_model`

Imputation is also introduced in learning incomplete data [123]. Dick et al. [34] present learning decision functions from the density mass of the distribution on finitely many imputations. Madsen et al. [86] implement the idea of multiple imputations from Bayesian statistics. They learn a Point Distribution Model (PDM) [31] from dataset containing only incomplete shapes and a full shape template, by first estimating the posterior distribution of point-to-point registrations for each partial observation. Then, they construct the PDM from the set of registration distributions. In our method of learning a full statistical spine model, only partial observations are present in the training set. We use PPCA to learn a statistical full spine surface model. PPCA has been used for craniofacial statistical modeling [81]. However, in their work full observations of the skull were used for training. In contrast, our proposed model never observed a complete spine.

2.2.2 STATISTICAL SPINE MODEL

The study of the human spine statistical shape model has attracted research attention for its many potential applications, such as image segmentation, bio-mechanics or pathology detection. However, as of today there is no publicly available statistical model of the 3D surface of the full spine. This is mainly due to the lack of openly available 3D data where the full spine is imaged and segmented. Therefore, most work has been done to model the shape of the vertebrae which are consistently observed in the scans. For instance, lumbar vertebrae have received a special attention for low back pain as this area supports the greatest load of the spine [17]. Hollenbeck et al. [46] generated a statistical shape model of the lumbar region from 52 subjects using PCA with a focus on the L4-L5 and L5-S1 functional spinal units (FSU). Campbell et al. [19] used automated methods for landmark identification to create a statistical shape model of the lumbar spine, applied in bio-mechanics and population-based finite element modeling. A non-linear SSM based kernel PCA was investigated for 3D active shape model segmentation [61]. They used 7 CT scans covering vertebrae level from T10 to L3.

Another difficulty in learning the statistical models of the spine is that the relative position of vertebrae highly depends on the patient's posture during data acquisition. Overlapping and excessive separation of neighboring structures may appear in the vertebrae segmentation. In order to tackle this problem, a model of the interspace of the vertebrae was developed [22], by considering the variations of the space between two surfaces. Similarly, our model only focuses on the relative translation of the vertebrae and disregards their relative rotation. Multi-vertebrae statistical models [15, 97] capture the variations in shape and pose simultaneously, and reduce the number of registration parameters, allowing to help in the segmentation of a section of the vertebral column. These models were tested on 32 subjects and with a focus on the lumbar vertebrae section.

While most works have focused on the modeling of a specific part of the spine, a few works have modeled the full spine. For instance, Klinder et al. [62] designed an automatic framework segmenting vertebrae from arbitrary CT images with a full spine model. To create the model they first scanned a commercially available plastic phantom to create the template, and then they manually registered it to 10 actual scans of the full spine.

Mirzaalian et al. [89] learned a statistical shape model of the spine by independently learning three models. one for each level (cervical, thoracic and lumbar). Thus, their per-level models do not learn the shape correlations across the full spine.

Other probabilistic models, different from the shape surface models, such as probabilistic atlas [99], graph models [100], hidden markov models [41] and hierarchical models [18, 101, 124] have also been proposed. For example, Ruiz et al. [99] proposed a probabilistic atlas of the spine. By co-registering 21 CT scans, a probability map is created which can be used to segment and detect the vertebrae with a special focus on ribs suppression. Schmidt et al. [100] proposed a probabilistic graphical model for the location and identification of the vertebrae in MR images. In both cases, full spines were observed at train time and the proposed methods cannot be used to infer the shape of the full spine from a partial observation.

2.3 LEARNING A STATISTICAL FULL SPINE MODEL FROM PARTIAL OBSERVATIONS

The inputs to our method are i) 80 manually segmented CT volumes ($\boldsymbol{V}^i$, $i \in [1, 80]$) with the annotations of the individual vertebra from VerSe'19 [102] and ii) 24 artist created mesh templates, one for each individual vertebra. The template meshes are symmetric with respect to the medial plane and we name them $\boldsymbol{T}_k$ with $k \in [1, .., 24]$ being the vertebra index and i the volume index. Each template $\boldsymbol{T}_k$ has N_k 3D vertices. From the segmented volumes we extract meshes using the Marching Cubes approach [79]. We name them scans and note them as $\mathcal{S}_k^i$. Let us note that: i) in the VerSe'19 Dataset none of the volumes contains the 24 vertebrae; ii) some vertebrae are imaged at the boundary of the CT volume and thus only provide an incomplete observation. These scans have holes and are not watertight; we note them $\bar{\mathcal{S}}_k^i$. An overview of the VerSe'19 Dataset vertebrae is shown in Fig. 2.1.

The creation of the statistical spine shape model consists of several steps: 1) Initial individual vertebra non-rigid registration; 2) PCA guided individual vertebra registration; 3) Learning PPCA on the registration; 4) Spine model creation. To compute the registrations we optimize objective functions with the dogleg gradient descent method [120]. We compute the derivatives using the auto-differentiation tool Chumpy [78].

2.3.1 INITIAL REGISTRATION

First, we perform a non-rigid registration of each template $\boldsymbol{T}_k$ to all watertight scans $\{\mathcal{S}_k^i - \bar{\mathcal{S}}_k^i\}$, by optimizing the point to surface distances between the scan and the template meshes. Because the scans are watertight, we want all template vertices to be close to the scan surface, and vice-versa, all scan vertices to be close to the template surface. We effectively enforce this constraint by computing the vertex to surface distance from a point set to a surface and define the energy $\mathrm{E}_{\mathrm{p2m}}(\mathcal{S}, \boldsymbol{T})$, which accounts for the distance of the

vertices of $\mathcal{S}$ to the mesh surface $\boldsymbol{T}$. The registration has three steps. Translation optimization, translation and rotation optimization and free form optimization. To simplify notation we drop the indices k and i.

We start by computing a translation $\boldsymbol{t}$, so that $\boldsymbol{T} + t$ is close to $\mathcal{S}$ by minimizing

$$\mathrm{E}(\mathcal{S}, \boldsymbol{T}, \boldsymbol{t}) = E_{p2m}(\mathcal{S}, \boldsymbol{T} + \boldsymbol{t}) + \lambda_{m2s} E_{p2m}(\boldsymbol{T} + \boldsymbol{t}, \mathcal{S}) \tag{2.1}$$

w.r.t. $\boldsymbol{t}$, where $\lambda_{m2s} = 1$. Next we optimize for a 3D rotation, parametrized by a 3D Rodriguez vector $\boldsymbol{r}$. Given a 3D vector $\boldsymbol{r}$ and a mesh $\boldsymbol{T}$, we use $R(\boldsymbol{T}, \boldsymbol{r})$ to describe the 3D rotated mesh. The CT scans from the VerSe'19 Dataset have a consistent patient orientation encoded in the dicom metadata. After reorienting all the CTs along the same orientation, we can thus initialize the rotation vector r ($r = [-2, 0, 2]$) with the same value for all $i \in [1, 80]$, and optimize

$$\mathrm{E}(\mathcal{S}, \boldsymbol{T}, \boldsymbol{t}, \boldsymbol{r}) = E_{p2m}(\mathcal{S}, R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}) + \lambda_{m2s} E_{p2m}(R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}, \mathcal{S}) \tag{2.2}$$

w.r.t. $\boldsymbol{t}$ and $\boldsymbol{r}$, where $\boldsymbol{t}$ is initialized with the result of (2.1). Next, we allow all vertices of $R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}$ to freely deform to best match $\mathcal{S}_k^i$. This free-form deformation is represented as an additive 3D vector added to each vertex that we note $\boldsymbol{f}$. To regularize the position of these displacements we use a coupling term on edges $\mathrm{E}_{cpl}(\boldsymbol{T}, \boldsymbol{T} + \boldsymbol{f})$, enforcing the edges of the registration to be close to the edges of the initial template shape. To compute E_{cpl} we use the same energy term defined in Eq. 8 from [14] and optimize

$$\begin{aligned}\mathrm{E}(\mathcal{S}, \boldsymbol{T}, \boldsymbol{t}, \boldsymbol{r}, \boldsymbol{f}) = E_{p2m}(\mathcal{S}, (R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}) + \boldsymbol{f}) + \\ + \lambda_{m2s} E_{p2m}((R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}) + \boldsymbol{f}, \mathcal{S}) + \\ + \lambda_{cpl} \mathrm{E}_{cpl}((R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}), (R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t}) + \boldsymbol{f})\end{aligned} \tag{2.3}$$

wrt to $\boldsymbol{f}$ and by keeping $\boldsymbol{t}$ and $\boldsymbol{r}$ fixed. The weight $\lambda_{cpl} = 0.01$ for all experiments. To bring all registrations to a coherent frame, we compute a rigid transformation [8] between the registration and the template. We name these coherent registrations $\mathcal{I}_k^i$.

2.3.2 PCA GUIDED REGISTRATION

For each vertebra k, we compute Principal Component Analysis (PCA) on all the available registrations. As we have few observations, we exploit the medial plane symmetry: for each registration we create its symmetric registration. With the registrations and their symmetry we obtain a (symmetric) mean shape $\boldsymbol{T}_k^{\mu}$ and it's principal vectors $\boldsymbol{B}_k$.

With the obtained PCA, we perform a second non-rigid registration of each template $\boldsymbol{T}_k$. This time all scans $\mathcal{S}_k^i$, including the non-watertight, are registered. We initialize the template k vertices with the mean shape $\boldsymbol{T}_k^{\mu}$ and use the computed shape space $\boldsymbol{B}_k$ to constrain the registration process. We compute the rigid transformation to obtain the coherent registrations and keep the corresponding rotation R_k^i and translation t_k^i.

In the PCA-guided registration, we first compute $\boldsymbol{t}$ and $\boldsymbol{r}$ for non-watertight scans $\bar{\mathcal{S}}_k^i$, by optimizing (2.1) and (2.2). We set $\lambda_{m2s} = 0$ as not every vertex in the template must explain a scan point.

Then, we use the shape space to parametrize the meshes as $\boldsymbol{T}^\mu + \boldsymbol{\beta}\boldsymbol{B}$ where $\boldsymbol{\beta}$ is a 10-dimensional vector and $\boldsymbol{B}$ the 10 first PCA principal components. We solve for $\boldsymbol{t}$, $\boldsymbol{r}$ and $\boldsymbol{\beta}$ to minimize

$$\mathrm{E}(\mathcal{S}, \boldsymbol{T}^\mu, \boldsymbol{t}, \boldsymbol{r}, \boldsymbol{\beta}) = E_{p2m}(\mathcal{S}, R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t} + \boldsymbol{\beta}\boldsymbol{B}) + \\ + \lambda_{m2s} E_{p2m}(R(\boldsymbol{T}, \boldsymbol{r}) + \boldsymbol{t} + \boldsymbol{\beta}\boldsymbol{B}, \mathcal{S}) \quad (2.4)$$

with $\lambda_{m2s} = 1$ if the scan is watertight and $\lambda_{m2s} = 0$ otherwise.

Finally we perform a free vertices optimization following (2.3). This time we couple the vertices to the solution obtained from (2.4). Then, we compute the rigid transformation [8] between the obtained registration and the mean template $\boldsymbol{T}^\mu$. We exploit again the medial plane symmetry and obtain a new set of registrations $\mathcal{A}_k^i$. From these registrations we learn the full-spine shape model.

2.3.3 PPCA ON THE REGISTRATION

For each volume index i, we do not have the full set of registration $\mathcal{A}_k^i$, as some k are not observed in the volume i (see Fig. 2.1). Thus we use Probabilistic Principal Component Analysis (PPCA) [110], a variant of PCA dealing with missing data. We construct a matrix with size $N \times S$ values, where $N = \sum_{k=1}^{24} N_k * 3$, and $S = 160$ is the number of volumes used (80 volumes plus their symmetric version). Each column i is the concatenation of the 24 vectors, each given by the registration $\mathcal{A}_k^i$ vertices, if it is available, or an empty vector if the data is missing. The created matrix has 57% missing values. We compute PPCA using the publicly available implementation in [115] and obtain the mean spine $\boldsymbol{T}^\mu$ and the associated shape variations $\boldsymbol{B}$.

2.3.4 SPINE MODEL

The format of the spine model is a graphical model mesh. It is parametrized by the shape parameters $\boldsymbol{\beta}$, and the pose parameters $\boldsymbol{\theta}$ and $\boldsymbol{t}$ where each of the 24 vertebrae has its own translation and rotation. The shape parameters $\boldsymbol{\beta}$ apply additive offsets to each vertebrae mesh. The offsets are the PCA principal components learned with the PPCA and the poses and translations are rigidly applied to each vertebra. To constrain the locations of the vertebrae, we learn a distribution over the relative positions of neighboring vertebrae. For each pair of neighboring vertebrae k and $k+1$, we use R_k^i, t_k^i and $t_{(k+1)}^i$ to compute the relative translation of vertebra $k+1$ w.r.t. vertebra k for the volume i. Then for each pair $k, k+1$ we fit a Gaussian model to the set of observed relative translations and obtain 23 distributions. As the relative rotation between vertebrae highly depends on the patient posture we do not attempt to learn a pose prior on the relative rotations.

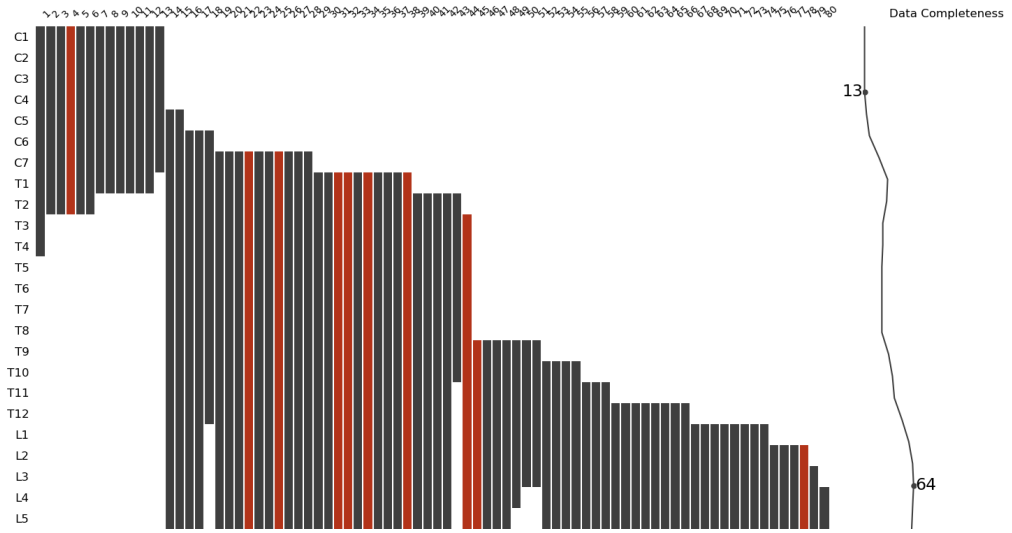


Figure 2.1: Overview of the VerSe'19 challenge dataset. Each row represents one of the 24 individual vertebrae, from C1 to L5. Each column is an observed patient. No full spine is observed for any patient. The columns in red are the test set of one of the 8-fold used for cross validation.

2.4 EXPERIMENTS

First of all, we introduce the dataset we use. Then to demonstrate the accuracy and benefits from the spine model, we perform two kinds of experiments. In a first set of experiments we assess the metric accuracy of the created model and the compacity of the learned shape space. In a second set of experiments we assess the accuracy of our model to predict the shape of missing vertebrae.

2.4.1 DATA PREPARATION

To train our model we use the publicly available VerSe'19 dataset [76][102] from the MICCAI Challenge 2019. The dataset provides 80 CT scans with manually annotated voxel-wise labels, and 40 CT scans without labels. We use the 80 subjects with labels to create the spine model. An overview of the available observations from the full spine in the Verse challenge datset is shown in Fig. 2.1. Each row in the figure represents a vertebrae, from top to bottom from C1 to T5. Each column represents a CT volume. As we can see, each CT scan only covers a few vertebrae. No full spine is observed in any CT volume. In addition, the number of volumes observing the given vertebrae is not balanced across different vertebrae. The dataset is dominated by lumbar observations and few cervical vertebrae are observed. The section observing the junction between cervical and thoracic vertebrae is critically sparse.

	C1	C2	C3	C4	C5	C6
	0.20,0.16	0.20,0.16	0.19,0.15	0.19,0.15	0.19,0.15	0.18,0.15
	C7	T1	T2	T3	T4	T5
mean, std (mm)	0.21,0.16	0.21,0.17	0.21,0.17	0.21,0.16	0.21,0.16	0.21,0.17
	T6	T7	T8	T9	T10	T11
	0.21,0.17	0.22,0.19	0.23,0.19	0.23,0.20	0.23,0.20	0.25,0.22
	T12	L1	L2	L3	L4	L5
	0.25,0.21	0.25,0.21	0.26,0.22	0.27,0.23	0.28,0.25	0.29,0.26

Table 2.1: Accuracy of individual vertebra registrations: point to mesh distance (mean, std) mm between each scan and its registration.

In our work, we aim at learning the shape correlations across the full spine. We use cross validation to evaluate our model. In order to make sure that the junction between the cervical and thoracic vertebrae is always seen at train time, we label a volume *junction volume* if C5 to T2 are observed in the volume, and *not junction volume* otherwise. We use this junction label to create a balanced cross validation 8-fold setting.

As our method is trained on very few samples, we carefully inspected the input data. We noticed some misannotations in the ground truth labels. They fall into two categories. One group is the true negative labels, appearing as floating points far away from the corresponding vertebra, such as lumbar volumes having cervical labels. To deal with these misannotations, we cleaned the ground truth meshes by keeping only the biggest connected component in the mesh extracted from the CT volume. The other group of misannotations is the false negative labels, in the form of real vertebrae regions not labeled and considered as background. We left these misannotations untouched and did not exploit the vertebrae information in these volumes.

2.4.2 Model accuracy

To assess the metric accuracy of the proposed model, we first measure how well the individual vertebrae registrations match the input data. Then we measure the compacity and the generalization power of the full-spine model.

Individual vertebrae registrations accuracy To assess the accuracy of our individual registrations we compute the point to mesh distance between the scans $\mathcal{S}_k^i$ and their registrations $\mathcal{A}_k^i$. For each individual vertebra we aggregate the mean distance per scan by taking the mean over all available volumes. We report the mean and standard deviation (mean, std) for each vertebrae in Tab. 2.1.

To aggregate the data, we do a weighted mean, by weighting each vertebrae errors with the number of samples and obtained 0.24 mm errors in mean and 0.20 mm std. The registrations faithfully capture the input data shape with a sub-millimeter precision.

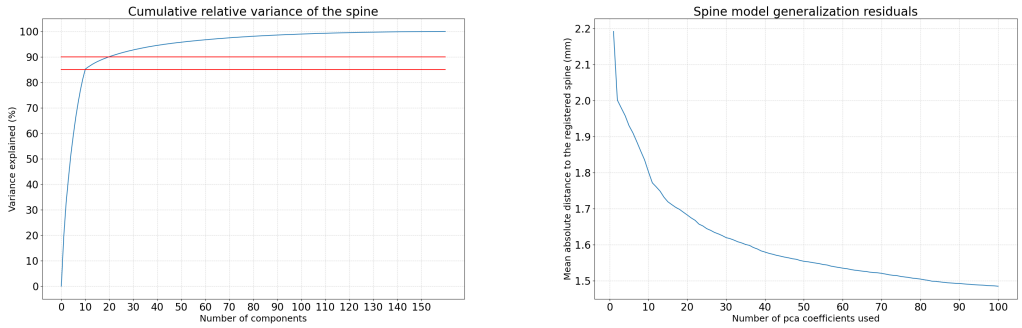


Figure 2.2: Left: the cumulated variance of the full-spine model w.r.t. the number of used shape components. Right: the generalization error of the model (cross validation) w.r.t. to the number of used shape components.

Full-Spine model accuracy. To assess the compacity of the learned spine shape space we show on the left of Fig. 2.2 the cumulative variance of the shape space. The first 10 shape components capture 85% of the variance, and 90% variance is captured by the first 20. To assess how well the model can capture the shape of unseen spines, we perform an 8-fold cross validation (see Sec. 2.4.1). We learn the shape space with a train set and evaluate it on the unseen test set. For an unseen and potentially partial registration, we complete the missing vertebrae by using the learned PPCA. We then project it into the PCA space and reconstruct it with a selected number of components. We then measure the vertex to vertex error between the registration and the reconstruction. Only original data, i.e, not completed by the PPCA, is used in the metrics. On the right of Fig. 2.2 the generalization errors w.r.t. the number of used components are presented. Using 10 components we obtain a mean reconstruction error of 1.8 mm, and using 20 components we obtain a mean error of 1.6 mm.

2.4.3 Prediction of missing vertebrae shape accuracy

In this experiment we mimic the scenario where the centering of the patient at CT acquisition time was inaccurate and some desired vertebrae were not imaged. Given a partial observation of the spine, with the learned model, we can reconstruct the missing spine shape. We set up the following experiment to measure the accuracy of the shape of the predicted vertebrae.

For each cross validation fold, we learn the spine model on the train est. Then we take all registrations in the test fold containing the cervical vertebrae C1 to C7. We first mask one vertebra starting from C1 and reconstruct it using the rest of the volume vertebrae. We then remove two vertebrae (C1 and C2) and reconstruct them using the rest. Progressively we remove cervical vertebrae until C5, and use the rest for reconstruction. We measure the vertex to vertex distance between the reconstructed vertebrae and the original registration. In an analogous way we also perform this experimental for lumbar vertebrae, by progressively removing L5 to L1.

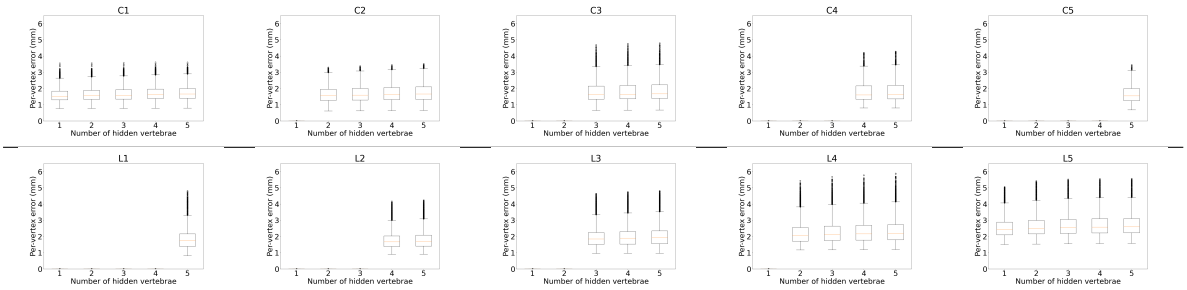


Figure 2.3: First and second rows: reconstruction errors of missing cervical vertebrae. Second Row: reconstruction errors of missing lumbar vertebrae. Refer to the text for the experimental setting description.

We use the 8-fold cross validation scheme (see Sec. 2.4.1) and evaluate on all volumes for which the data is available. In Fig. 2.3 we report the per vertex mean error over all volumes. In the first row, the reconstruction errors of the cervicals are shown. The most left plot reports the reconstruction errors (RE) on C1, when masking 6 vertebrae. The second plot, RE on C2, when masking 5 vertebrae. The third plot RE on C3, when masking 4 vertebrae, the forth plot RE on C4, when masking 3 vertebrae, and the fifth plot, RE on C5 when masking 2 vertebrae. In the second row, the results for the analogous lumbar experiment are presented. In Fig. 2.4 we present the per vertex mean errors on the mean template.

We aggregate the per vertex errors for all C1 and obtain reconstruction errors of 1.61 mm when one vertebra is masked, 1.65 mm when two vertebrae are masked, 1.68 mm when three vertebrae are masked, 1.72 mm when four vertebrae are masked, 1.75 mm when five vertebrae are masked. For the L5 vertebra we obtain 2.57 mm when one vertebra is masked, 2.65 mm when two vertebrae are masked, 2.70 mm when three vertebrae are masked, 2.73 mm when four vertebrae are masked, 2.74 mm when five vertebrae are masked.

A trend in the increase of the error is observed as more vertebrae are masked, but the errors do not increase drastically. This trend can be seen in Fig. 2.3 as well as in the aggregated errors. These experiments show that the shape of all vertebrae is strongly correlated throughout the spine. By observing a subset, a plausible shape of the following vertebrae can be inferred.

2.5 CONCLUSION

In this chapter we present a statistical surface model of the full-spine learned solely from partial and incomplete views of vertebrae. In order to deal with the partial observations we use probabilistic principal component analysis (PPCA) to learn a surface shape model of the full spine. Quantitatively evaluation demonstrates that the obtained model faithfully captures the shape of the population in a low dimensional space and generalizes to left out data. The model is made available to the community for non-commercial purposes at `https://gitlab.inria.fr/spine/spine_model`.

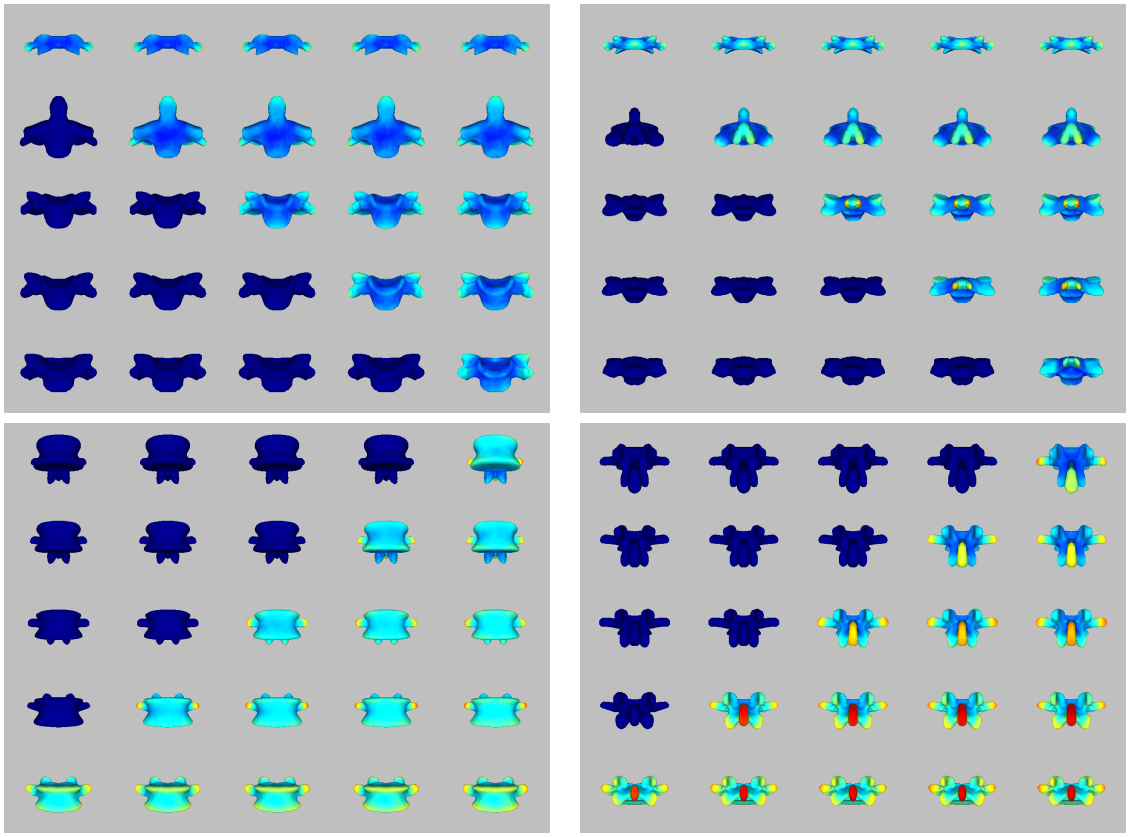


Figure 2.4: First Row: per vertex mean reconstruction error of missing cervical vertebrae visualized on the mean vertebra. Second Row: per vertex mean reconstruction error of missing lumbar vertebrae visualized on the mean vertebra. Red is 6 mm. Dark blue is used when the vertebra is not masked. Left: front view. Right: back view

We presented an application of the model to predict the shape of missing vertebrae. The model can also be applied in other tasks, such as the generation of synthetic data, its registration to 2D Xray images and its use in the segmentation of volumetric images.

CHAPTER 3

SPINE SEGMENTATION

3.1 INTRODUCTION

For diagnosis and treatment of the spine related diseases, imaging is required. Computed tomography (CT) and magnetic resonance (MR) imaging are two main technologies that used in the computed aided spinal intervention and surgery. MR imaging is widely recognized as the imaging technique of choice for accessing the inter-vertebral disc degeneration and abnormalities due to its excellent soft tissue contrast [37]. While CT is broadly used in the examination of vertebral fractures and spinal curvatures as it presents high contrast for bony structures. To prevent damage to nearby blood vessels and nerves during the spine surgery, the vertebrae and their surrounding tissue must be precisely localized. Thus, pixel-wise segmentation of the spine and individual vertebrae is essential.

An optimal solution is to develop a method to output a multi-label segmentation mask where individual vertebrae are identified and segmented. It is challenging as the CT is usually with arbitrary field of views, varying number of vertebrae and their similar shapes make the final output spine not well sorted. Besides, the spatial consistency of each individual vertebra mask is difficult to be maintained when performing the multi-label segmentation. Therefore, most methods develop multi-label segmentation networks assuming a specific area (eg. lumbar vertebrae) is observed [97, 53]. Multi-step methods are also exploited. For the multi-label segmentation, it will be comprehensively introduced in Chapter 5.

In fact, a segmentation task is a pixel-wise classification task. Compared with classifying each pixel in CT image into one of the multiple classes, classifying each pixel into two classes is much easier as the intra-variance is largely reduced. Namely binary segmentation. In this chapter, we introduce a spine binary segmentation approach. The spine binary segmentation serves as a substantial component in multi-stage vertebrae segmentation and labelling. It not only indicates where all individual vertebrae locate, but also describes the morphology of the vertebrae. Obtaining an adequate spine binary segmentation is a basic step for the following investigations of individual vertebrae identification and segmentation.

3.2 Related Work

For the spine and vertebrae segmentation, earlier methods use primarily statistical models [62, 97], active shape models [32, 11, 4, 33], deformable models [68, 67, 59, 83], graph cuts [10], or level sets [112].

Rasoulian et al. [97] use a statistical multi-vertebra model that treats the shape and pose of each vertebra independently. They also propose a novel iterative expectation maximization registration technique to align the multi-vertebral model to CT spine images. Howe et al. [47] describe a semi-automatic segmentation method for application to cervical and lumbar X-ray images. The coarse-to-fine method consists of three stages where the segmentation process utilizing the generalized Hough transform for one stage, and active appearance models for two stages. Registration is also used in segmenting the vertebrae by propagating the manual annotation through different modalities of the same subject.

Model-based segmentation usually requires a complete vertebra presented and visible in the image which is not robust to arbitrary field of view. However the deformable model tends to relax the constraints while keeping the shape priors. Korez et al. [68, 67] propose an approach based on interpolation theory to locate the whole spine in 3D CT scans and detect individual vertebrae locations from the whole spine. The vertebrae are further segmented by a shape-constrained deformable approach. Similar approach of adapting shape-constrained deformable models to individual vertebrae is also performed by Klinder et al. [62]. For the individual vertebrae locations, they first adapt tube-shape segments to extract the spine curve then detect the individual vertebrae on curved-planar reformatted images using the generalized Hough transform. Ma et al. [83] introduce a hierachical coarse-to-fine deformable surface-based segmentation that relied on response maps of a trained bone structure edge detection algorithm. Kim et al. [59] use a deformable model that is based on constructing 3D fences to separate vertebrae from vally-emphasized Gaussian images, and then the region growing algorithm is applied within the constructed 3D fences to obtain the final segmentation.

Level set is employed in vertebrae segmentation framework by Lim et al. [73], they extract local geometrical features using the Willmore flow and prior shape knowledge using kernel density estimation. Khandelwal et al. [57] also use geometrical features that they propose to segment the spine from CT images employing geometric flows, especially the use of anisotropic diffusion filtering combined with flux maximizing flows. They use shape prior based method to semi-automatically segment individual vertebra by manually put the seed at the center of the vertebral body. Hammernik et al. [44] introduce a total variation based framework that incorporates a prior model of a vertebral mean shape, image intensity and edge information.

Other shape representations are also used in the segmentation task. Ibragimov et al. [50] present a segmentation framework, in which a novel landmark-based shape representation of vertebrae is combined with game-theoretic landmark detection augmented by the strategy dominance concept. Kadoury et al. [54, 55] build an articulated shape manifold

by embedding the vertebrae from a database into a low-dimensional sub-space, and apply the Marcov random field optimization to infer between the shape manifold and the shape of unseen vertebra.

As we see, for the multi-vertebrae segmentation, existing methods usually locate the individual vertebrae first using a coarse-to-fine strategy from the spine to each vertebra, then perform the segmentation by registering the shape model to individual vertebrae. It is powerful as they encode the prior shape knowledge. However, the shape constraints of surface models or image appearance are not robust to the pathology, i.e. vertebral compression, fractures. Nevertheless, the coarse-to-fine individual vertebra localization turns out to be a stable and accurate strategy to deal with the vertebrae similarities while separating them. The spatial consistency is maintained when segmenting a target single vertebra.

Recent learning-based approaches have nevertheless demonstrated promising performance. Some feature-based methods treat all image voxels belonging to the observed vertebra as points of interest, and perform the segmentation using the intensity-based features by machine learning classification techniques. Ghosh et al. [40] propose a fully automated method for segmentation of lumbar vertebrae from clinical CT volumes based on geometry and intensity features. The inter-vertebral discs are located and labeled using a probabilistic model leveraging the pixel-level and object-level features and the vertebral bodies are segmented using a series of morphological and hole-filling operation. Huang et al. [48] use an iterative normalized-cut segmentation algorithm to segment the precise vertebra regions from the detected vertebrae locations. An optional case-adaptive segmentation approach is proposed by Kelm et al. [56], that allows to segment the spinal disks and vertebrae in MR and CT respectively.

The development of deep neural networks even boost the spine and vertebrae segmentation performance to a higher level. Al et al. [5, 6] introduce a novel shape-aware term in the loss function of a deep segmentation network which learns to segment cervical vertebrae from X-rays images while preserving the shape of the target object. The shape constraint is encoded in the form of vertebra mask contour. The method is semi-automatic for the vertebra center point is manually provided at test time. They further improve the work by predicting the shape directly. Instead of performing the segmentation as pixel-wise classification, they predict the shape as signed distance function. A novel loss function that computes the error directly in the shape domain in contrast to the other deep networks where errors are computed in a pixel-wise manner. Sekuboyina et al. [104] propose a two-stage approach for lumbar vertebrae segmentation in which they first regress the bounding box of the lumbar region and then segment and label each vertebra locally. Pang et al. [92] propose a detection-guided mixed-supervised segmentation network, which consists of a segmentation path for generating the spine mask and a detection path for regressing heatmaps prediction of keypoints. Both vertebrae and discs are segmented while the lumbar area is treated. Several methods achieve very good performances by assuming that a specific part of the spine (usually the cervical or lumbar level) is observed [53, 6, 7]. However, these methods are not robust to arbitrary field of views. Lessmann et al. [70] handle this issue by iteratively applying convolutional networks

in the CT images. The vertebrae are segmented and identified as they are progressively found with a sliding window.

Whitehead et al. [118] propose a series of four pixel-wise segmentation networks which segment vertebrae and discs from MR images at different scales. Results from one network in the chain are fed as input to the next network in the chain. Both the original image and the results from previous network are used as input to the next network, producing an increasingly refined segmentation result. To enhance the boundary information of the segmentation, Vania et al. [114] develop a method utilizing convolutional neural networks and class redundancy as a soft constraint to segment the spine from CT images. Specifically, besides the label for the spine and background, they create two redundant classes for the bone boundary to emphasize the contours. Kolařík et al. [64] introduce a 3D Dense-U-Net architecture for high resolution 3D volumetric segmentation. They pretrain a denoising network to initialize the segmentation network. Dense-U-Net is also used in [27] where they use a two-stage Dense-U-Net for vertebrae localization and segmentation. They first use a 2D-Dense-U-Net to localize vertebrae with dense labels on 2D slices. Then they use a 3D-Dense-U-Net to segment specific vertebra within a region-of-interest based on the detected centroid.

Multi-stage method is widely employed for segmenting multiple vertebrae. Masuzawa et al. [87] propose a multi-stage framework where, first, the bounding boxes of cervical, thoracic and lumbar vertebrae are found, then the vertebrae in each bounding box are segmented and identified in an iterative manner. Payer et al. [94] use a three-step approach to first localize the spine, then simultaneously locate and identify each vertebra and segment each individual vertebra within a window. A similar approach is found in [58] to detect and segment vertebrae in X-ray images.

3.3 SPINE BINARY SEGMENTATION

The input to our method is a CT volume with arbitrary field of view, arbitrary resolution and arbitrary anatomic position, potentially imaging the spine of a human. First the given CT volume is automatically sampled into an isotropic resolution of $1 \times 1 \times 1\ mm^3$ and oriented to *PIR* anatomic orientation, where P stands for *towards the back of the body*, I for *towards the bottom of the body* and R for *radial aspect of the forearm*.

3.3.1 NETWORK ARCHITECTURE

From the pre-processed CT volume, we compute a binary segmentation using an Attention U-net [91] deep neural network. The backbone architecture is a 3D U-net [98] with attention blocks embedded in the skip connections. Traditional skip connections propagate the information from the compression path to the reconstruction path. The attention gate enables to highlight the salient features (corresponding to bone) passed to reconstruction path in order to suppress irrelevant regions (such as soft tissue) in the input image. The output of the network is a probability mask (with floating values in range [0, 1]) for each voxel indicating if it belongs to the spine or not.

An illustration of the network architecture and the attention gate is shown in Figure 3.1. We name one *convolutional layer* as a 3D convolutional layer followed by a BatchNormalization layer [51] and a ReLU activation function. The 3D convolutional layer is of kernel size 3, padding 1 and stride 1. The encoder consists of 5 blocks. The first block is constructed with one padding layer and two *convolutional layers* which outputs a feature map with the same size as input image. The $2nd$ to $5th$ blocks contains one 3D Max Pooling layer of stride 2 and two *convolutional layers* respectively. The size of the output feature maps is downsampled by a factor of 2. The number of features passed through the 5 blocks are $16, 32, 64, 128, 256$. Similarly, the decoder consists of 5 blocks. Each block contains two *convolutional layers*, one upsamping layer of scale factor of 2 and two *convolutional layers*. The input of each block is the concatenation of the output from the previous block and the output of the corresponding encoder layer through the attention gate (see Right in Figure 3.1). The number of features though the decoder blocks are $256, 128, 64, 32, 16$. The output from the decoder is further passed through a single 3D convolutional layer of kernel size 1. The output of the network maintains the same size as the input image.

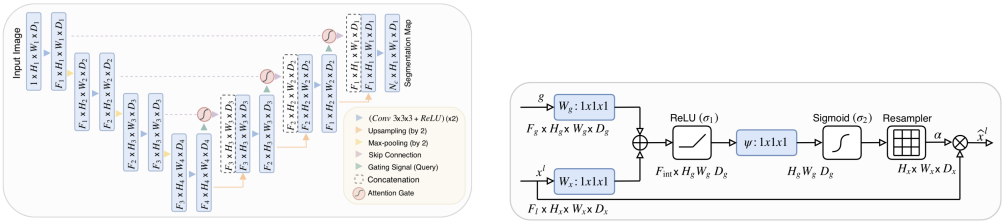


Figure 3.1: Left: architecture of the attention U-net network. Right: iIllustration of the attention gate. The illustrations are from [91].

As the input CT has an arbitrary size and the network accepts only a fixed sized input, we use patches of size $96 \times 96 \times 96$ as the input to the network. At inference time, the model is applied in a sliding mode over the whole CT volume with a stride length of 24 (one quarter of the patch size) and for each voxel we obtain 64 predictions. We aggregate the results for each voxel by averaging the 64 probability maps and binarize the result with a 0.5 threshold.

3.3.2 Loss functions

To train the network we use the regression loss which is the addition of Dice coefficient and Mean Square Error (MSE), the objective being to minimize the difference between our prediction and the ground truth. Dice coefficient is a similarity index measuring the spacial overlap of two sets of data. It is widely used in medical image segmentation to address the data imbalance problem [126]. Mean Square Error calculates the Euclidean distance between two sets of data and average the sum of the distances. We utilize these two metrics in our loss function to minimize the dissimilarity from both global and local aspects. Formally, we note the input as x, the output as $G(x)$, and the ground truth patch

y and minimize

$$\mathbf{L}(G(x), y) = \mathbb{E}_{x,y}[1 - \frac{2|G(x) \cdot y|}{|G(x)|^2 + |y|^2}] + \lambda \mathbb{E}_{x,y}[\|y - G(x)\|_2], \tag{3.1}$$

where $\lambda = 10$ was empirically set and kept constant in all our training processes.

3.3.3 IMPLEMENTATION DETAILS

Data augmentation Although the CT input is in *PIR* anatomic orientation, some vertebrae especially cervical vertebrae have an important rotation around Z-axis. In addition, patients have slightly variant rotations with respect to the scanning device. To make the network more robust to these cases, we augment the data with rotations in $[-30, 30]$ degrees around the Z-Axis, the rotation is w.r.t. the axial plane. We then generate the training data by cropping the augmented volumes with a sliding window of size $96 \times 96 \times 96$ and with a stride of 90. As the background has much larger proportion than the foreground (spine) which introduces data imbalance issue. We then generate additional training data by only cropping the central area of CT volume where the spine lies, with a stride of 30. In total 48853 training cubes are generated from 80 CT scans.

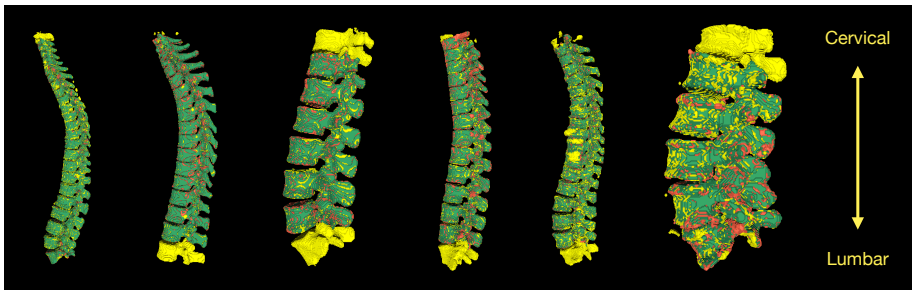
Training The method is implemented using Python 3.7 and PyTorch 1.4. A batch size of 10 is used for parallel computing on two GPUs of Quadro RTX 5000. We use Adam optimization [60] with a learning rate of $1e-3$. The training runs for 200 epochs and we select the model which has the best performance on validation set.

3.3.4 EVALUATION

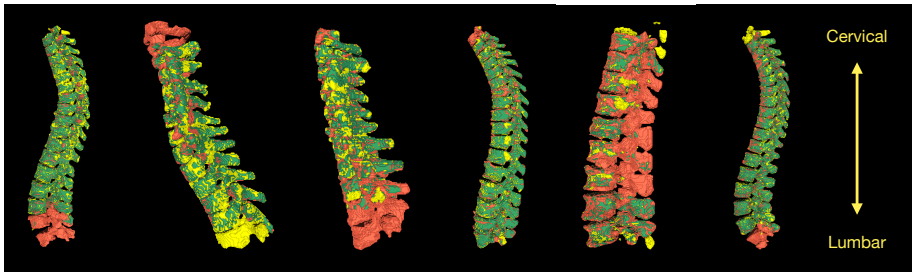
We use VerSe20 challenge dataset [102] to develop the method and evaluate the results. The public training data which contains 100 CT subjects is randomly split into a training set with 80 CT scans and a validation set with 20 CTs. The evaluation is on the public testset and hidden testset which has 100 CT subjects respectively.

Overall evaluation To evaluate the spine binary segmentation, we use Dice score (DSC) and Hausdorff Distance Error (HD). Dice score is a variant form of intersection of union to gauge the similarity of two segmentations. The Hausdorff distance measures how far two subsets of a metric space are from each other. It quantifies the largest segmentation error between two segmentation surfaces. We report the Dice score and Hausdorff distance error of the public testset and hidden testset in Table. 3.1.

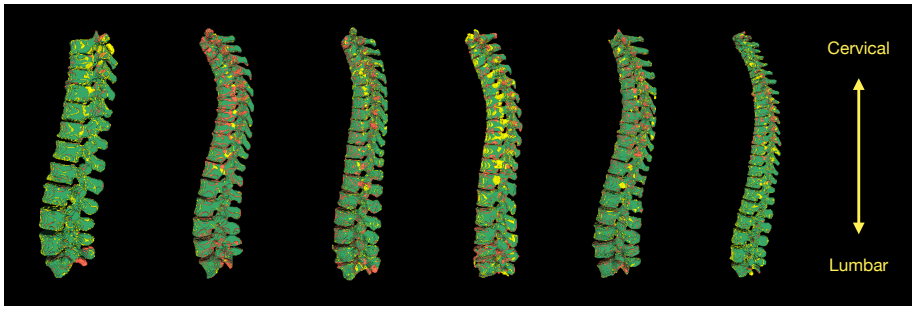
We obtain mean and median Dice score over 90% showing that our predicted spine binary segmentation is plausible with respect to the ground truth segmentation. However, the mean and median Hausdorff distance errors are over 20mm which is beyond a noise level. To further inspect the error, we visualize the results with color encoded: True positive predictions in green, true negatives in red and false positives in yellow. Seen



(a) False positive errors in the prediction



(b) True negative errors in the prediction



(c) Regular predictions

Figure 3.2: Errors visualization of spine binary segmentation prediction.

from Figure 3.2 (a), we found that for most cases, our method predicts more vertebrae than the ground truth, at where close to the image borders. Nevertheless, some corner cases are missed as shown in Figure 3.2 (b). It makes sense that the Hausdorff distance error is around 20mm which is close to an averaged vertebra length along the vertical axis. Figure 3.2 (c) represents the regular predictions where only surface errors are observed.

Per vertebra evaluation The spine binary segmentation indicates where all the vertebrae locate. It can be used as a reference for the further steps of individual vertebrae localization and segmentation. Thus, an optimal spine binary segmentation should cover all the individual vertebrae. The overall Dice score does not reveal this specification. To evaluate if the predicted spine binary segmentation covers all the individual vertebrae and how much they overlap, we propose a pseudo Dice score (pDSC) as metrics. Since the spine binary segmentation is united and there is no vertebrae boundary information, the

	DSC(%)↑		HD (mm)↓	
	mean± std	median	mean± std	median
public	94.51 ± 2.75	95.38	26.14 ± 22.84	23.34
hidden	94.34 ± 3.00	95.44	26.99 ± 16.16	22.82

Table 3.1: Evaluation of the spine binary segmentation on VerSe20 public and hidden testsets.

	public testset		hidden testset	
	mean± std	median	mean± std	median
unet	94.35± 2.70	95.36	**94.35± 2.78**	95.35
attention-net	**94.51± 2.75**	**95.38**	94.34± 3.00	**95.44**

Table 3.2: Dice score (%) evaluation of the Unet with and without attention mechanism.

pseudo Dice score depicts how much ground truth segmentation per vertebra has overlap with our predicted spine binary segmentation.

$$pDSC = \frac{2 \times G_i \cdot (G_i \cap P_s)}{{G_i}^2 + (G_i \cap P_s)^2} \tag{3.2}$$

G_i is the ground truth binary mask of individual vertebrae, P_s is the intersection of our predicted spine binary segmentation with the ground truth individual vertebra mask. The output is the percentage of every ground truth vertebra covered by our predicted spine binary segmentation.

In the public testset we observe that 94.65% vertebrae (1292 out of 1365) have a pseudo Dice score over 95%; 98.68% vertebrae (1347 out of 1365) have a pseudo Dice score over 80% and 99.78% vertebrae (1362 out of 1365) have a pseudo Dice score over 50%. Only one vertebra has a pseudo Dice below 40% (19.49%). In the hidden testset, 93.84% vertebrae (1264 out of 1347) have a pseudo Dice score over 95%; 99.33% vertebrae (1338 out of 1347) have a pseudo Dice score over 80% and 99.55% vertebrae (1341 out of 1347) have a pseudo Dice score over 50%; Five vertebrae having pseudo Dice below 40%. The spine binary segmentation is adequate enough but still misses vertebrae.

3.3.5 Ablation study

We use the Attention U-net in our method which aims to bring more attention to the spinal structures in CT images and achieve higher accuracy on the spine segmentation. To assess the effect of the attention gate, we perform an experiment using the classical U-net without attention mechanism. Table. 3.2 and table. 3.3 show the overall evaluation of DSC and HD on the spine binary segmentation predictions. A marginal improvement is observed on most mean and median values and a significant improvement is found on the HD error for public testset. Further we evaluate the performance on per-vertebra segmentation.

	public testset		hidden testset	
	mean± std	median	mean± std	median
unet	31.06± 51.45	23.68	**26.24± 18.18**	23.83
attention-net	**26.14± 22.84**	**23.34**	26.99± 16.16	**22.82**

Table 3.3: HD (mm) evaluation of the unet with and without attention mechanism.

For public testset, there are four vertebrae having a pseudo Dice score below 40% while there is only one when using attention gates. For hidden testset, 93.02% vertebrae have a pseudo Dice score over 95%. The improvement using attention gates (93.84%) is minor, but 12 more vertebrae are obtained when considering the count (1252 out of 1347). Same number of vertebrae (5) is observed having pseudo Dice below 40%.

3.4 SPINE MULTI-LEVEL SEGMENTATION

3.4.1 MOTIVATION

The spine binary segmentation is a union of all the individual vertebrae, while the optimal goal is to separate the neighboring vertebrae. An intuitive way is to perform the same strategy as spine binary segmentation but, instead of classifying each voxel into one of two classes (bone or not bone), multi-label segmentation classifies each voxel into one of multiple classes. It is more challenging as it potentially consists of two tasks, one is to recognize if it belongs to the spine or background, the other is to classify which vertebra it belongs to if it is not from the background. In this section, we adopt the same approach as spine binary segmentation to perform a multi-vertebrae segmentation. Rather than a united spine binary mask, we output both the individual vertebra mask and their labels. We explore 24-vertebrae segmentation, and 3-level segmentation which is associated with the cervical, thoracic and lumbar vertebrae. The aim is to investigate how well or worse the results can achieve and conceptualize the challenges of the multi-vertebrae segmentation using the same single network which provides promising results for binary segmentation.

3.4.2 METHOD

The method takes a 3D CT volume as input and output a multi-label segmentation mask. As stated in Sec 3.3, the CT volume is firstly pre-processed into the same resolution and orientation, then being segmented by an Attention-Unet in a sliding mode across the whole volume. The input to the network is of size $96 \times 96 \times 96$. The output of the network is a probability map of the same size as input with $N+1$ channels, where N represents the number of classes excluding the background. To train the semantic segmentation network, we employ both a regression loss and a classification loss which is the sum of the Cross Entropy and Mean Square Error (MSE). The objective is to maximize the probability of

Dice score (%)	public testset	hidden testset
24-vertebrae	44.94±14.54	44.92±15.35
3-level	87.04±6.17	87.74±5.98

Table 3.4: Quantitative results of the 24-vertebrae segmentation and the 3-level segmentation.

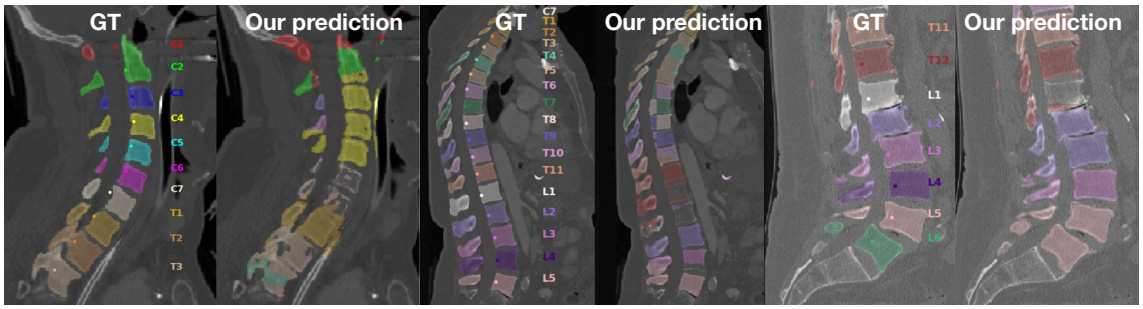


Figure 3.3: Results visualization of 24-vertebrae segmentation.

our prediction being correctly classified. Formally, we note the input as x, the output as $H(x)$, the ground truth patch y and minimize

$$\mathbf{L}(H(x), y) = \mathbb{E}_{x,y}[-\frac{1}{N+1}\sum_{i=1}^{N+1} \omega_i * y_i \log(H(x_i))] + \lambda \mathbb{E}_{x,y}[\|y - H(x)\|_2], \quad (3.3)$$

where $\lambda = 10$ was empirically set and kept constant in all our training processes. The weight ω is introduced to balance each class in order to have equal contributions to the learning regardless the imbalance of the training data.

3.4.3 Evaluation

We train two segmentation models, one performs 24-vertebrae segmentation ($N = 24$) and the other performs 3-level segmentation ($N = 3$). The models are trained with the VerSe20 [102] training set using the same train/validation split as in Sec 3.3.4. We evaluate the method on VerSe20 public and hidden test sets.

For the 24-vertebrae segmentation, only an averaged 45% accuracy is obtained as shown in Table 3.4. By inspecting the results (Fig 3.3), we found severe issues about multi-vertebrae segmentation using patches: a) the mask of individual vertebra is not spatially consistent, in the sense that a single vertebra contains more than one label; b) neighboring vertebrae are assigned with the same label, making them not separated; c) some vertebrae are not segmented or they are classified as background causing the obtained spine mask is not continuous. A strong argument towards these problems is the limited receptive field. The input to the network is always a fixed size patch which is much smaller than the entire CT volume. The convolutional network each time processes one patch, the convolved features cover only a relevant input image region. Therefore, the

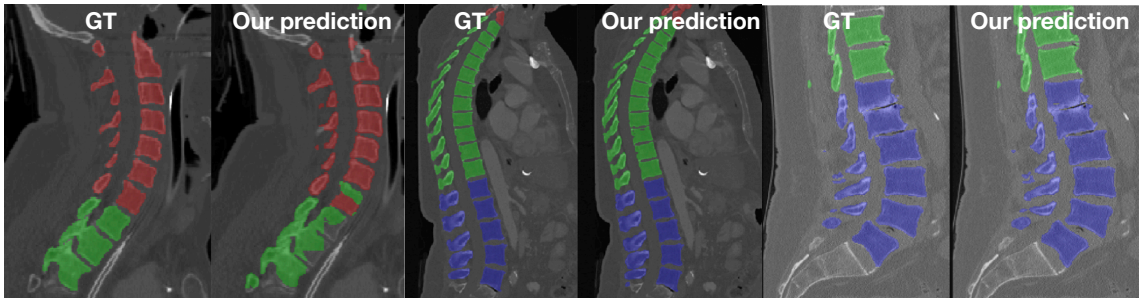


Figure 3.4: Results visualization of 3-spine-level segmentation. Cervical vertebrae in red, thoracic vertebrae in green and lumbar vertebrae in blue.

Dice score (%)	cervical	thoracic	lumbar
public testset	74.43±10.97	87.23±8.40	91.29±5.72
hidden testset	76.56±10.62	87.52±7.15	91.85±5.69

Table 3.5: Quantitative evaluation of 3-level segmentation results on each level.

correlation between the patches is missing. Though the neighboring patches have some overlapping areas, the consistency of the whole image are not taken into account. A way to optimize the multi-vertebrae segmentation from arbitrary field-of-view CT images using one network is to take as large receptive field as possible, ideally the whole image. In this way, the global information is received and the consistency of the vertebrae can be learned. However, as we know, CT scans are usually of large sizes. A single CT volume can contain up to millions of voxels. Taking a whole CT volume as input to the network is extremely computationally expensive. Thus, previous works were encouraged to use 2D slice or multi-stage approaches.

For segmenting the spine into cervical, thoracic and lumbar regions, we obtain an average of 87% Dice score (Table 3.4). The performance is better in terms of the mask completeness comparing with the 24-vertebrae segmentation, as noticed in Figure 3.4. The errors are most frequently observed in the transition of the spine levels. We further evaluate the performance of each spine level, the Dice scores are shown in Table 3.5. An averaged Dice score of 75%, 87% and 91% are obtained on cervical, thoracic and lumbar vertebrae respectively. An increment is observed from top to down. The performance of each level is approximately proportional to the ratio of the amount of each level of vertebrae in the training set, although they were weighed (Eq 3.3) during the training. Compared with the direct 24-vertebrae segmentation, segmenting the vertebrae into anatomical groups eases the challenges from classifying a voxel into one of 24 classes to one of 3 classes. The intra-variance is largely reduced, so that a higher accuracy is observed.

Moreover, we perform another experiment on evaluating the spine segmentation ability of the multi-vertebrae segmentations. Despite the classification capability, precisely segmenting the whole spine out of the CT scan is important and can be used as a support-

Dice score (%)	binary segmentation	3-level segmentation	24-vertebrae segmentation
public testset	94.51±2.75	90.48±4.13	78.20±8.76
hidden testset	94.34±3.00	90.11±4.55	77.45±7.76

Table 3.6: Quantitative evaluation of the spine binary mask by multi-vertebrae segmentation.

ive reference. We binarize the results of 3-level and 24-vertebrae segmentation to evaluate them as spine binary masks. As shown in Table 3.6, 3-level segmentation achieved around 90% Dice score and 24-vertebrae segmentation obtained about 78% Dice score. They are both lower than the pure spine binary segmentation (94%, Sec 3.3). Although the multi-vertebrae segmentation can perform the identification and segmentation of the vertebrae simultaneously, the accuracy is more difficult to be maintained as it does more than one tasks. This observation encourages us to explore multi-stage methods of segmenting, identifying and localizing vertebrae. Also, the experiment of 3-level segmentation inspires us to follow a hierarchical approach of identifying the vertebrae, which will be demonstrated in Chapter 4.

3.5 CONCLUSION

In this chapter, we thoroughly reviewed the previous work of spine segmentation. We proposed to segment the spine as a binary mask using Attention-Unet, promising results were obtained. We further explored the multi-vertebrae segmentation. The results in health is hardly about pleasant but it encouraged us in two aspects: i) to explore multi-stage method for an accurate vertebrae segmentation, localization and identification, while the spine binary segmentation can serve as an adequate reference of where the vertebrae locate and can be used for the latter steps of individual vertebra segmentation and identification; ii) a hierarchical strategy of vertebrae identification which they can be categorized into anatomic groups first and then into individual labels.

CHAPTER 4

VERTEBRAE IDENTIFICATION

4.1 INTRODUCTION

The human spine is usually composed of 24 vertebrae. They are structured in three anatomic groups: 7 cervical (C1-C7), 12 thoracic (T1-T12) and 5 lumbar vertebrae (L1-L5). Each group shares morphological and functional characteristics. For example, cervical vertebrae take control of the neck flexibility so that they are relatively in smaller size. Lumbar vertebrae supports the greatest load of the spine, they have the largest size among the three groups. Thoracic vertebrae connect to the ribs and can be used as a reference of locating other organs. Figure 4.1 illustrates the shapes of the three anatomic groups. Cervical vertebrae, especially C1 (Atlas) and C2 (Axis) have a rather distinctive shape. Other vertebrae, such as neighboring thoracic or lumbar vertebrae share a visually similar morphological appearance.

Automatic identification of vertebrae in spinal imaging, such as Computed Tomography (CT) or Magnetic Resource Imaging (MRI), is crucial in the context of clinical diagnosis and surgical planning. It is a key enabling component and a prerequisite for the downstream applications. However, there are critical challenges preventing a precise-enough automated identification scheme for clinical practice: i) Neighboring vertebrae can have similar shape which makes it difficult to robustly distinguish one from another. ii) The number of vertebrae can change among patients. While most humans have 12 thoracic vertebrae, some present 11 or 13. Similarly, some humans have 6 lumbar vertebrae instead of 5. These anatomically rare cases make the identification particularly challenging. iii) Pathological spines, e.g. suffering from bone calcification or broken vertebrae, can present abnormal shapes that affect the learned decision function. iv) Given arbitrary field of view CT images, counting vertebrae is suboptimal [70] when no distinctive vertebra is available in the observation which, again, impacts the identification.

As of today, the identification of all vertebrae in a CT volume still remains a major challenge for the community. Our method is designed to deal with these challenges by combing a local approach, aiming at identifying individual vertebrae, with a global

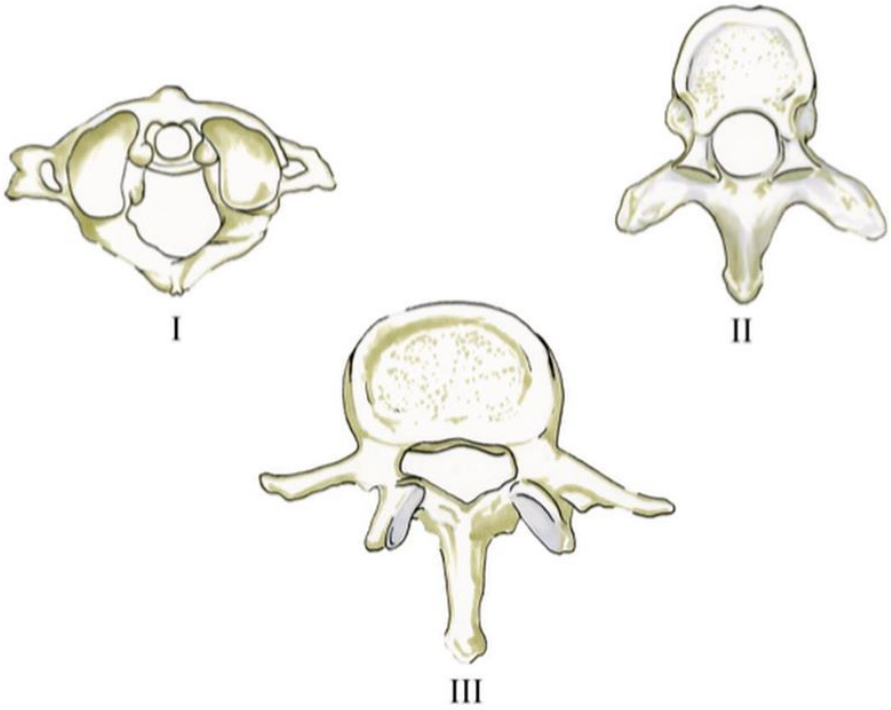


Figure 4.1: Illustration of the shape of the vertebrae. [1] I - Representative shape of cervical vertebrae C3-C7. II - Representative shape of the thoracic T1-T12. III - Representative shape of the lumbar L1-L5.

reasoning enforcing the anatomic consistency among the individual predictions. To be specific, instead of taking as input the raw data from CT images, we extract the shape of individual vertebra and quantify to which extent the shape of a single vertebra is discriminating. We experimentally quantify that solely using the shape of the vertebrae achieves higher accuracy than using as input the CT image. Provided a source of relevant probabilities of individual vertebra predictions, a graph is introduced to enforce the global consistency of the predicted individual vertebrae labels. It enables to handle anatomically rare cases, e.g. the transitional vertebrae T13, L6 and the absence of T12.

4.2 RELATED WORK

As vertebrae localization and identification are always seen as one task called vertebrae labeling, we review the vertebrae labeling literature, addressing both the localization and identification tasks. For vertebrae labeling, traditional methods build on shape models [69, 18, 74, 83, 66, 77] or graphical models [41, 16] to detect landmarks. Such models are powerful since they encode prior knowledge about the full structure of the spine. Ma et al. [83] carry out the vertebrae identification process by testing which mean shape has the maximum response. It requires to feed all the mean shapes to all the vertebrae present in the volume. In [41] vertebrae locations are regressed by making assumptions on the shape appearance, but the approach is not robust to pathological cases or abnormal images with

surgical implants. They further improve the method by transforming the sparse location annotation into dense labels in the surrounding area, which avoids modeling the vertebra shape [42]. Both approaches [41, 42] use random forests. Chen et al. [25] share a similar idea and localize the coarse vertebrae candidates by training a binary RF classifier based on HOG features. Huang et al. [48] propose a statistical learning approach based on an improved AdaBoost algorithm for vertebrae detection. Kelm et al. [56] use a generative anatomical network that incorporates relative pose information to simultaneously locate and label the spinal disks. An iterative version of marginal space learning is used to generate candidates using a learned prior on the individual nine dimensional transformation spaces.

More recent methods build on machine learning and convolutional neural networks (CNN) in which hand-crafted features tend to be replaced by learned ones. Sekuboyina et al. [103] combine information across several 2D projections using a butterfly-like architecture and encode the local spine structure as an anatomic prior with an energy-based adversarial training. Liao et al. [72] and Qin et al. [95] develop a multi-label classification and localization network using FCN and residual blocks. They improve the classification branch with bidirectional recurrent neural network (Bi-RNN) to encode the short and long range spatial and contextual information. In [117] a nnUnet keypoint detection model is trained to estimate 26 vertebrae (including the sacrum) activation maps. Chen et al. [26] propose a variant FCN that localizes vertebrae in original resolution and classifies them in down-sampled resolutions. These are stand-alone networks that directly output vertebrae locations and identifications and can be trained in an end-to-end manner. However, these methods struggle to handle rare cases (the missing T12 or the occurrence of T13 or L6) as the anatomic consistency is not explicitly enforced. In addition, the scarcity of data with rare cases makes the end-to-end learning approaches under-perform in their detection.

Other works in the literature perform vertebrae labelling in multiple stages. McCouat and Glocker [88] employ separate CNNs to localize the vertebrae in 3D samples and identify them in 2D slices with a two-stage method. In order to find vertebrae, Jakubicek et al. [52] detect firstly the spine and then track the spinal cord based on the combination of a CNN and a growing sphere method. Payer et al. [94] follow a similar strategy of coarse to fine vertebrae localization by detecting the spine in lower resolutions and each vertebra in higher resolution images. For identifying vertebrae, Chen et al. [25] propose a joint learning model (J-CNN) that can classify the vertebrae labels and encode the pairwise conditional dependency at the same time. In our work, we also use multiple stages to obtain the vertebrae locations and identifications. In addition, we loop through a consistency check in order to enforce mutual consistency.

Since raw predictions from networks are not always accurate, various post-processing stages have been adopted to constrain vertebrae locations based on the anatomical ordering. Huang et al. [48] refine the detected vertebrae locations by curve fitting, it removes the false detections and recovers missing detected vertebrae. Chen et al. [25] use a shape regression model to correct the offsets from the deviation in the vertical axis. The model assumes that the coordinates distribution can be described by a quadratic form and is limited to coordinates on the vertical axis. Yang et al. [121] introduce a chain-structure

graphical model to depict the spatial relationship between vertebrae and regularize their locations with an L_1 norm to learn the best sparse representation. Chen et al. [26] model the score maps interpolated along the 1-D spinal curve with a Hidden Markov Model to generate the optimized 1-D coordinates. Mader et al. [84, 85] learn an optimal conditional random field as a spatial regularizer to select a global optimal configuration over all the landmarks. In our approach, we propose an efficient graph that enforces inter-vertebrae constraints over the individual vertebra classification predictions, with the ability to model patients with or without transitional vertebrae.

4.3 INDIVIDUAL VERTEBRAE CLASSIFICATION

In this section, we aim at exploring if individual vertebra is morphologically distinguishable and quantifying to which extent the shape of a single vertebra is discriminating. This knowledge can provide a relevant information to the global full spine identification methods. For example, distinctive vertebrae (such as C1, C2) can be used as reliable anchor points in global configuration, whereas the contribution of unreliable predictions can be diminished.

To this end, we propose to train a 3D convolutional neural network (CNN) that, given the morphology of an isolated vertebra as input, one can predict the probability of the vertebra belonging to one of the 24 classes. Then, motivated by the fact that the anatomic groups have distinctive features, as illustrated in Fig. 4.1, we study a second approach by first identifying which anatomic group (cervical, thoracic, lumbar) the vertebra belongs to, and then its individual identification with a per-group specialized network. We further quantify that by including the contextual information, higher accuracy can be achieved for individual predictions. To perform our study, we use the publicly available dataset VerSe'19 [102]. As the medical imaging dataset usually contans a small number of annotated samples, we study the impact of several augmentation techniques (translation, rotation, noise addition) in the classification task. Preliminary results show that: i) the shape of an individual vertebra can be used to faithfully identify its anatomic group (cervical, thoracic, lumbar), ii) the shape of the thoracic vertebrae appears to have the highest similarity and are the ones where the network is confused the most. iii) including specially neighboring vertebrae can boost the individual vertebra prediction accuracy.

4.3.1 DATA EXTRACTION

The dataset used for this study is VerSe'19 [102], which is a spine dataset with 80 CT scans annotated. The vertebrae are manually labeled and segmented in the form of a binary mask. We use the ground truth (GT) segmentation masks of the vertebrae as input data to train our classifier. We extract every individual vertebra from the GT annotation and obtain a binary volume for each vertebra.

Re-sampling and re-orientation The CT scans are with arbitrary field of view, resolution and anatomic position. We pre-process the volumes by first re-sampling them into

an isotropic resolution of $1 \times 1 \times 1$ mm^3 and re-orientating them into a PIR anatomic orientation, where P stands for *towards the back of the body*, I for *towards the bottom of the body* and R for *radial aspect of the forearm*.

Connected components As the dataset is annotated by hand, the segmentation masks contain some noise in the form of small isolated clusters of voxels. After extracting the individual vertebrae, we use 3D connected component algorithm to extract the largest connected component to obtain a desired and clean vertebra, discarding the floating and unconnected voxels.

Padding Volumetric networks take as input a predefined volume structure having the same size. However, the extracted individual vertebrae have different dimensions as they are in different sizes. To decide on the network input size we use the largest vertebra dimension in the dataset plus a margin of approx. 20%, namely a cube of size $128 \times 128 \times 128$. All binary volumes are then centered and zero-padded to match the cube size.

4.3.2 Data augmentation

A common useful technique to overcome the scarcity of data while training neural networks is data augmentation. Proposing variations of the same instance is essential to lean a robust model. That is, adding more samples to leverage the training process by learning more complex features that consolidate the network's discriminative ability towards more robust features. We use four augmentation techniques: rotations, translations, scaling and additive noise.

We consider the rotations around the three axes, hence introducing multiple orientations to the network. Because the acquired volumes have a coherent global orientation (PIR), we uniformly sample angles θ in the range of $[-20, 20]^3$ degrees.

In order to introduce the translation invariance we feed as training samples translated inputs of the desired objects. So, instead of learning a centered vertebra in the cube, samples are shifted by a $\delta \in R^3$ offset. The translation is uniformly sampled from the interval $[-20, 20]^3$ mm.

Another transformation we consider is the scaling property of an object. We apply a uniform scaling factor $\gamma \in R$, uniformly sampling from the interval $[0.8, 1.2]$.

Additive noise is also added to the input data. We use salt-and-pepper noise by sampling from a Poisson distribution with parameter $\lambda = 0.05$.

Our mechanism to generate an augmented dataset involves applying 10 random combinations of the aforementioned transformations to each vertebra. Hence, 10 random versions of each vertebra are used as training samples. All these transformations are conducted while preserving the input cube size as $128 \times 128 \times 128$.

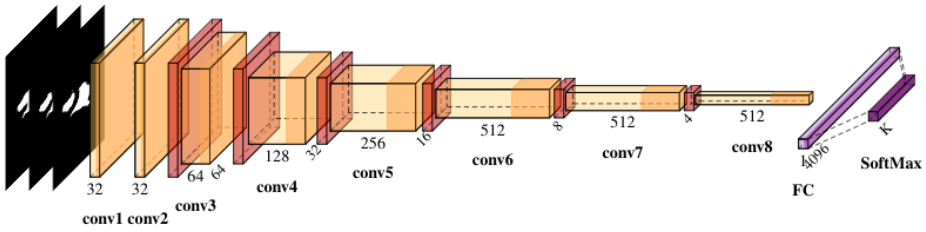


Figure 4.2: Individual vertebra classification CNN architecture. The numbers under each block describe the size of the output kernels after each operation. The numbers on the z-axis of each block describe the size of the output cube in R^3.

4.3.3 NETWORK ARCHITECTURE

The input to the network is a binary volume of size $128 \times 128 \times 128$ that contains a vertebra segmentation mask. The output is the N dimension vector of all classes probabilities. The CNN architecture is composed of 8 convolutional blocks followed by a fully connected layer of size 4096, illustrated in Fig. 4.2. Then a SoftMax activation layer is applied to get the probabilities of each class of the input volumes. Each convolutional block (except for the first 2 blocks) consists of 4 layers: a 3D downsampling pooling layer of kernal size $2 \times 2 \times 2$ that shrinks the input volume to half of its size; a 3D convolutional layer of kernel $3 \times 3 \times 3$ with stride 1 and padding 1; a Batch normalization layer [51] of momentum 0.95 for computing the running mean and variance and a ReLU non-linearity activation function.

To identify the vertebra, we propose three approaches:

1. Given an individual vertebra as input, the network predicts one out of the 24 classes ($N = 24$).

2. Given an individual vertebra as input, the network first predicts which anatomic group it belongs to ($N = 3$), then a group-specialized network is applied to predict the individual class ($N = 7/12/5$).

3. Given a vertebra with its neighboring vertebrae (with the target vertebra centered) as input, the network predicts its class using the two-stage method.

4.3.4 LOSS FUNCTION

We choose to use Cross Entropy loss to minimize the difference between our prediction and the ground truth,

$$\mathbf{L}(p(x), y) = -\frac{1}{N} \sum_{i=1}^{N} \omega_i * y_i \log(p(x_i)) \tag{4.1}$$

	Rot	Trans	Rot+Trans	All
train set	99.5	99.8	99.8	100
validation set	73	70	75	81
test set	74	67	70	80

Table 4.1: Ablation study on the impact of different data augmentation strategies. The classification accuracy of train, validation and tests set of one of the 8 folds is reported. *Rot* is only applied with rotation augmentation; *Trans* is only applied with translation augmentation; *Rot+Trans* combines random transformations for each strategy on each sample; *All* includes all previous augmentations with scaling and the noise addition. The highest accuracy is consistently obtained by the *All augmentation* strategy.

where y represents the ground truth, $p(x)$ represents the predicted probability. To solve the problem of class imbalance in the dataset, a weight ω is included. The weight is computed as inversely proportional to the numbers of each vertebra level.

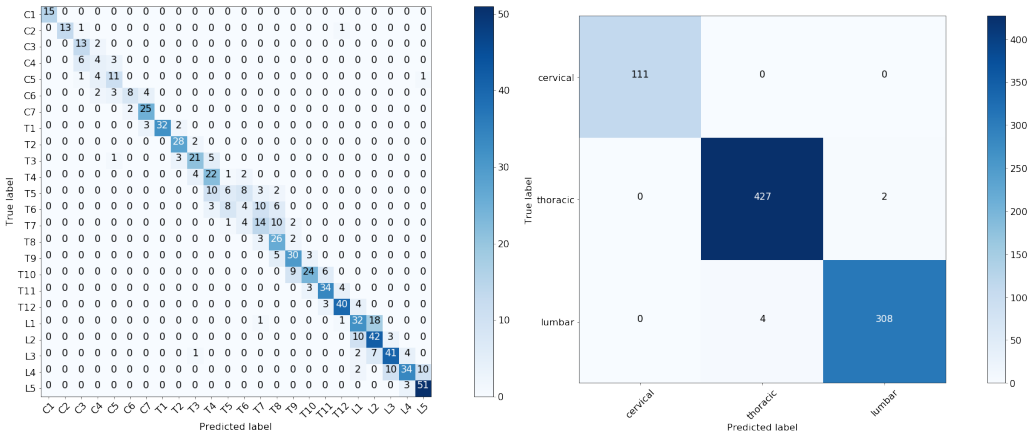
4.3.5 EXPERIMENTS

In this section we perform several experiments. First we evaluate the direct classification of a vertebra into its 24 classes possibilities. Then we evaluate the two-stage classification scheme, where the anatomic group is identified and the individual class is predicted by a group specialized network. Furthermore, we quantify that including contextual information boosts the classification accuracy. The dataset is randomly split into training (60/80), validation (10/80) and test set (10/80). We use a 8-fold cross validation to precisely test the model's ability.

Data augmentation study In order to access the relevance of different augmentation strategies, we perform an ablation study by each time applying one augmentation technique or the combinations. Table 4.1 presents the classification accuracy of each augmentation strategy and each set. The results show the benefit of the different augmentation strategies, with the *All* technique systematically obtaining the highest accuracy. Thus, we use *All augmentation* strategy for the following experiments.

24-level vs. two-stage classification We first train our model to classify, individually, directly into one of the 24 classes. The obtained result achieves 71% classification accuracy. To further inspect the results, we present the confusion matrix for the 24-level classification in Fig. 4.3. For the cervical and lumbar group, the only confusion arises with direct neighboring vertebrae. However, the vertebrae in the middle of the thoracic segment (T5 to T9) present the highest confusions.

Then we experiment if a single vertebra can be distinguished in terms of anatomic groups. Instead of predicting the individual label, we train a group classifier. For each vertebra, we predict its group label (cervical, thoracic, lumbar). The model achieves 99.3% accuracy with the confusion matrix presented in Fig. 4.3. The result shows that



(a) 24-level (b) 3-group

Figure 4.3: Confusion matrices of individual vertebra classification. (a) 24-level classification. (b) 3 anatomic group classification.

the morphology of an individual vertebra contains relevant information to accurately distinguish the anatomic group. It motivates us to study if a per-group classifier can better classify the individual vertebrae.

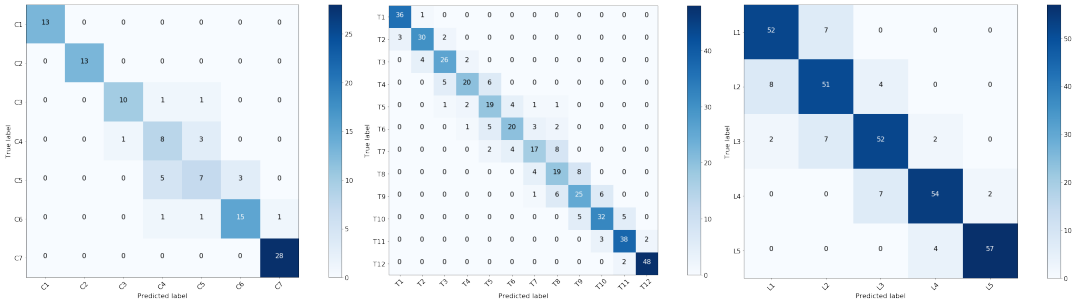
Knowing an individual vertebra is distinctive among the three anatomic groups. We then train three per-group models for the cervical (7 classes), thoracic (12 classes) and lumbar vertebrae (5 classes) respectively.

For the cervical group, we obtain an accuracy of 84.68%. Figure 4.4 shows the confusion matrix of the cervical classifier. As expected, C1 and C2 have a rather characteristic shape, making them easy for the network to be accurately identified. We observe that most confusing part happens between C4, C5 and C6, while C3 and C7 can be distinguishable with less confusions.

For the thoracic group, we obtain an accuracy of 76.92%. Figure 4.4 shows the confusion matrix of the thoracic classifier. Most confusions arise in the section between T5 and T9, where distant vertebrae up to two neighbors are wrongly predicted (T7 for T5, T8 for T6 or T7 for T9). It indicates that the shape of the middle section of the thoracic vertebrae is the most similar, making them less individually identifiable.

For the lumbar group, we obtain an accuracy of 86.08%. Figure 4.4 shows the confusion matrix of the lumbar classifier. It is worth noting that the failures are evenly distributed with the direct neighboring vertebrae. While L1 and L2 have a slightly higher confusion rate, the shape of L4 and L5 is seen to be more distinguishable with less confusions.

With vs. without contextual information We extend the two-stage method by adding the neighboring vertebrae segmentation masks to the input, while keeping the target vertebra in the center of the cube as shown in Fig. 4.5. One of the motivations of including



(a) Cervical group (b) Thoracic group (c) Lumbar group

Figure 4.4: Confusion matrices of group-specialized classification. (a) Cervical group. (b) Thoracic group. (c) Lumbar group.

(a) without context (b) with context

Figure 4.5: Input cube to the vertebrae classification network. (a) Individual vertebra. (b) Individual vertebra with neighboring context.

the contextual information is that the spine binary segmentation we obtained from Chapter 3 is united and the individual vertebra mask segmentation is not separated. Moreover, multi vertebrae includes the discs information which can be complementary and benefit the model learning. We quantify that by including the neighboring vertebrae segmentation, the classification accuracy increased from 83% to 88%. The overall comparison of the three strategies is present in Tab. 4.2.

4.4 GLOBAL VERTEBRAE IDENTIFICATION

From the previous section, we know that several vertebrae can be morphologically distinguishable while most vertebrae are possibly but not precisely recognized because of their similar shapes. To address this challenge, we propose to combine a local and a global reasoning. For the local reasoning, we use an individual classification method predicting one of the 24 possible labels for each vertebra. At a larger scale, we use the prior knowledge

	cervical	thoracic	lumbar	individual
strategy 1	71%	61%	75%	71%
strategy 2	85%	77%	86%	83%
strategy 3	**93%**	**81%**	**90%**	**88%**

Table 4.2: The accuracy of individual vertebra classification. Strategy 1: 24-level classification; Strategy 2: two-stage classification; Strategy 3: two-stage method with neighboring vertebrae as input. Strategy 3 consistently achieves highest accuracy.

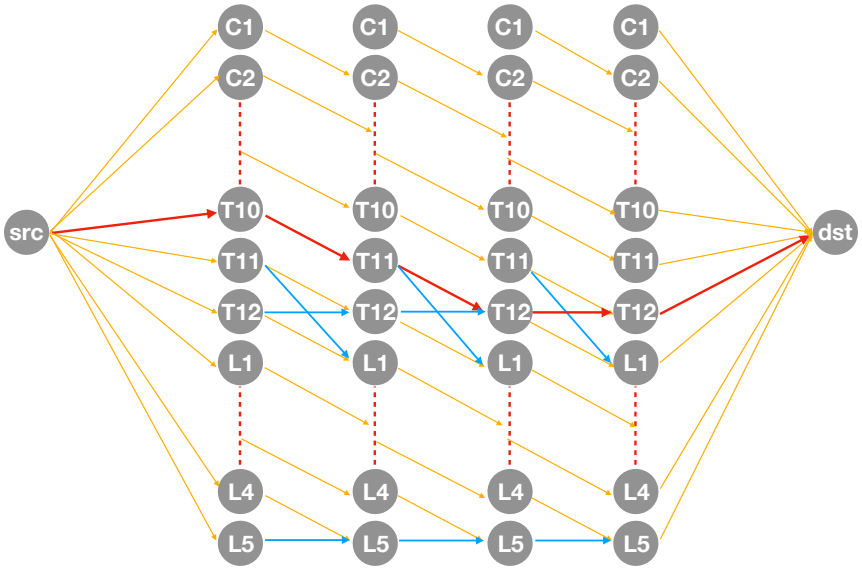


Figure 4.6: Spine identification global graph. Example with 4 vertebrae. Edges in orange are regular connections between nodes. Edges in blue connecting T12 to T12 (and L5 to L5) allow for the presence of T13 (and L6). The edge between T11 and L1 allows for T12 to be absent. The result with red edges corresponds to a spine where T10, T11, T12 and T13 are observed. (Best viewed in color)

that vertebrae are sorted in the spine and aggregate the local information into a global consistent final identification.

In order to deal with pathological cases, such as T13 or T6, for which there's only a small amount in the dataset. In the local individual vertebra classification, we consider T13 as T12 and L6 as L5 and assign their fine identification in the global reasoning.

4.4.1 Graph optimization

The local classification is not perfect and some vertebrae can be misidentified. However we propose to take advantage of a global approach to enforce the natural ordering of vertebrae, i.e. consecutive vertebrae should get consecutive labels. To enforce this global consistency we formulate the problem as a shortest path problem in a graph, as shown in Fig. 4.6. Given a list with n consecutive locations we create a graph with $n \times 24$

nodes. Each of the n columns represents a location, and each of the 24 rows represents a vertebra class. We formulate two types of cost, the unary and the binary costs. We populate the unary cost of each node with the cost obtained from the individual local probability, a value in the range [0,1]. Additionally, as the 3-group predictions are very reliable, we also add a cost to each unary to prevent group swapping. The binary costs are encoded in the edges of the graph. These edges between nodes encode the knowledge if two consecutive vertebrae have a consistent class or not, i.e., node N_i^j with $i \in [1, n]$ and $j \in [1, 24]$ is only connected with N_{i+1}^{j+1}. This effectively enforces that consecutive vertebrae get consecutive labels.

Three configurations need the special attention. They correspond to non standard configurations: i. the presence of T13, ii. the absence of T12 and iii. the presence of L6. To handle the presence of T13, an edge is added between two consecutive T12 nodes, shown in blue in Fig. 4.6. For the lack of T12, an edge is added between two consecutive nodes belonging to T11 and L1, shown in brown in Fig. 4.6. The presence of L6 is dealt with an edge between two consecutive L5 nodes, shown in blue in Fig. 4.6. These special edges are given a higher cost. To complete the graph, two extra nodes with no cost are added to be used as the source (*src*) and the destination (*dst*) in the optimization. The *src* node is connected to all nodes of the first vertebrae and the *dst* node is connected to all nodes of the last vertebrae.

Once the graph is created and the costs are populated, we compute the shortest path in the graph using the classical Dijsktra algorithm [35]. As a post process, we check if repeated instances of T12 (or L5) are obtained and adjust the class of the second node to T13 (or L6) accordingly.

4.4.2 Evaluation

To evaluate the full identification framework including individual vertebrae classification and the graph optimization, we use a larger dataset VerSe'20 [102]. It's distributed into public training set, public testset and hidden testset. Each set consists of 100 CT scans. We use the training set and the associated annotations to set up our method. 80 random CT scans from the training set are used for training the classification network, both the ground truth segmentation and our predicted segmentation (from Chap. 3) are given to train the network. The rest (20 CT scans) are used for validation. Then we evaluate the identification performance using our predicted segmentation of the public testset.

Figure 4.7 shows an overview of the public training set structure. Each column represents a scan of a patient, and each row represents the presence of one of the 26 vertebrae, from C1 to L6. The arbitrary field of view of the input images can be observed, in that each scan covers often a few vertebrae only. The data is class imbalanced: more thoracic and lumbar vertebrae are present in comparison to fewer cervical vertebrae. Let us note that atypical anatomies are present in this dataset: in the transitional vertebrae there are cases with occurrences of T13 or absence of T12, as well as some occurrence of L6 cases.

Table 4.3 presents the identification results on the public testset. Using the two-stage method (strategy 2), we obtain an accuracy of 70.50%. By considering the context of

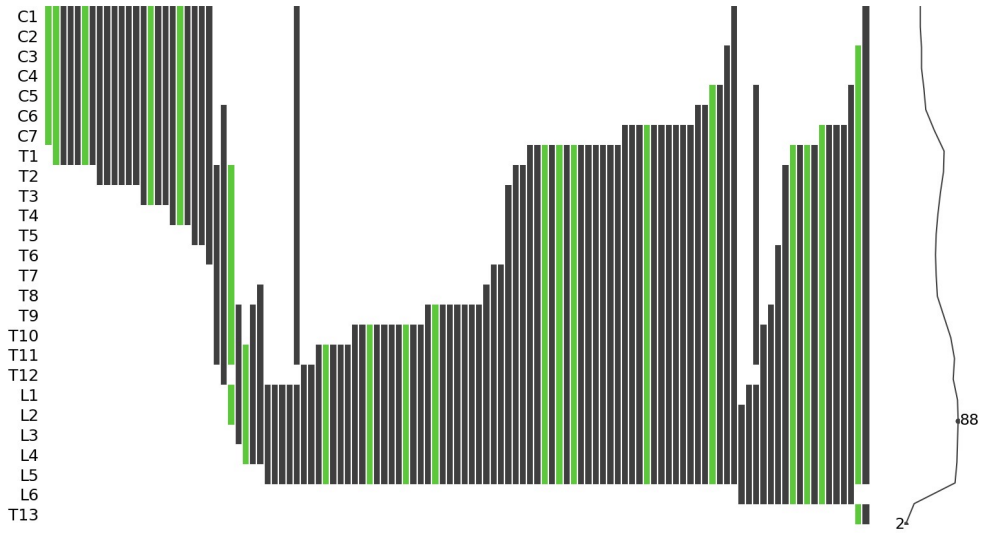


Figure 4.7: Dataset coverage (VerSe20 challenge public training set). Gray: our training set; Green: our validation set. Each column represents a scan of a patient, and each row represents the presence of one of the 26 vertebrae, from C1 to L6. On the right the aggregated number of vertebrae is shown. (Best viewed in color)

	group	cervical	thoracic	lumbar	individual
w/o	99.48%	82.32%	67.44%	72.85%	70.50%
w/	**99.56%**	**96.95%**	**83.25%**	**79.63%**	**84.55%**

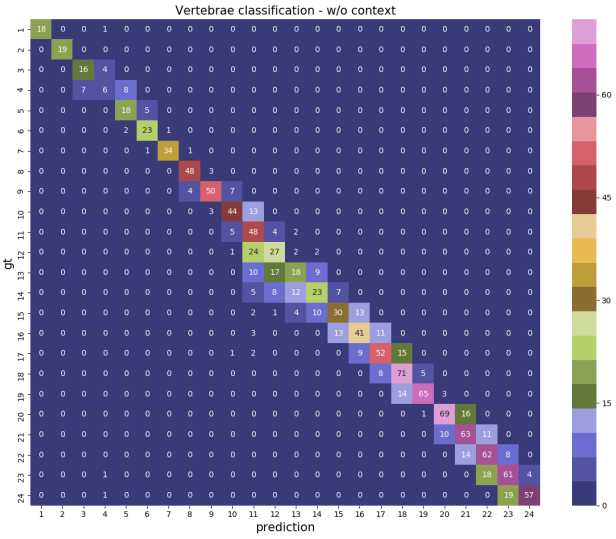
Table 4.3: Individual vertebrae classification accuracy with and without context.

the neighboring vertebrae, we achieve an accuracy of 84.55%, which represents an improvement of 14.05% with respect to the single vertebra approach. We also provide the confusion matrices in Fig. 4.8. It is worth noting that the thoracic and lumbar vertebrae are the ones with most confusions and the most common confusion is a prediction of the next or previous vertebra.

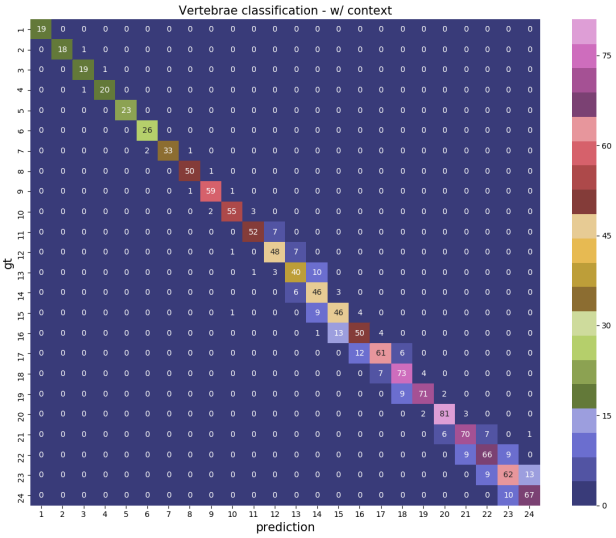
To evaluate the benefit of the proposed global individual identification, we performed the graph optimization with the obtained predictions from the local identification. The identification accuracy is improved from 84.55% to 97.36%. As the output of the graph is always a consistent spine, the only failures are the ones caused by a label shift (see Fig. 4.8).

4.5 Ablation Study

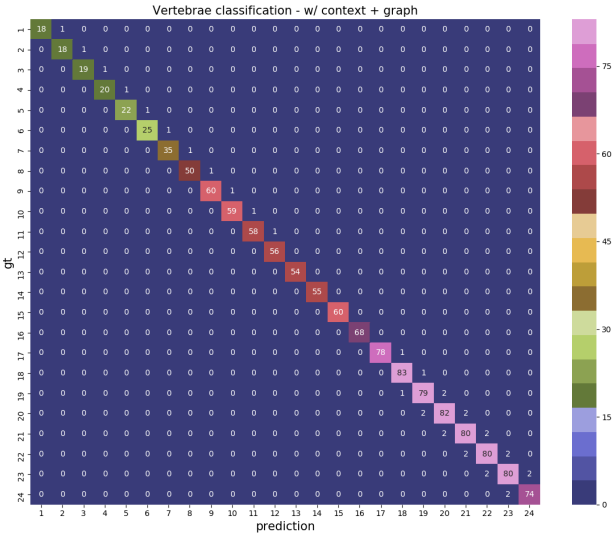
In our proposed approach, we study if an individual vertebra can be morphologically distinguishable and to what extent, where we took the shape of the vertebra as input to the classification network. However, most works proposed to identify the vertebrae from the



(a) w/o context



(b) w/ context



(c) w/ context+graph

Figure 4.8: Confusion matrices of Individual vertebrae classification.

	w/o graph	w/ graph
CT image cube	82.36%	96.27%
binary mask cube	**84.55%**	**97.36%**

Table 4.4: Vertebrae identification accuracy using CT image and binary mask as input, with and without graph optimization.

raw data, such as CT images [117] or MRI [119], as they hyperthesize the best information and context should come from the actual CT volume. We perform an ablation study to compare the performance of different inputs.

Instead of taking a binary volume consisting of a vertebra mask along with its neighboring vertebrae masks as input, given the vertebra location we crop a cube of size $128 \times 128 \times 128$ from the CT image where the vertebra is centered. To indicate the location of the target vertebra, we add a second channel that has the same size as the cube. The location channel contains the 3D location converted into a 3D probability map using a Gaussian kernel with $\delta = 20$. We then use the same two-stage method (Sec 4.3.3) which first predicts the anatomic group, then predicts the individual label for each input. We report the obtained accuracy in Tab. 4.4.

We notice that by using the original intensity data from the CT image, we obtain the identification accuracy of 82.36%, a bit lower than taking the segmentation of the vertebrae as input (84.55%). It indicates that the information which benefits the classification network mainly comes from the shape of the vertebrae. Segmenting the vertebrae from CT image helps reduce unnecessary information and highlight the shape features, which is a key component to tell apart the vertebrae.

4.6 CONCLUSION

In this chapter, we address the challenge of vertebrae identification. We propose an approach which combines a local and global reasoning. We first present a study on the classification of the vertebrae using its morphology. From the result of the experiments, we observe that the individual vertebrae classification is a non-trivial problem due to the morphological similarity of neighboring vertebrae. However, our experiments confirm that classifying a vertebra into its anatomic group is relatively straightforward for the networks. It inspires us to use a hierarchical two-stage method that first to predict the anatomic group a vertebra belongs to, then a group-specialized classifier is applied to predict the individual vertebra label. Further, we quantify that by including the context of neighboring vertebrae segmentation masks, a higher classification accuracy is achieved. Moreover, we also quantify that using the morphology of the vertebrae achieves higher accuracy than using the raw data from CT image. As the local classification is able to provide the probabilities of the vertebrae labels but they are not perfect, we then use the prior knowledge that vertebrae are sorted in the spine to construct a simple graph. The graph performs a global reasoning and select an optimal configuration of the vertebrae identifi-

cations. Moreover, it allows us to well handle the abnormal cases such as the presence of T13, L6 and the absence of T12.

CHAPTER 5

VERTEBRAE SEGMENTATION AND LABELLING

5.1 INTRODUCTION

Vertebrae localization, segmentation and identification in CT images is key to numerous clinical applications. In previous chapters, we propose to segment the spine out of the CT scan as a binary mask and identify each vertebra using their morphology and a graphical model given the vertebrae locations. The remaining task is that the vertebrae are not individually segmented and their locations are not provided. At the aim of outputing a multi-label vertebrae segmentation, while deep learning strategies have brought to this field significant improvements over recent years, transitional and pathological vertebrae are still plaguing most existing approaches as a consequence of their poor representation in training datasets. Alternatively, proposed non-learning based methods take benefit of prior knowledge to handle such particular cases. In this chapter, we propose to combine both strategies. To this purpose we introduce an iterative cycle in which individual vertebrae are recursively localized, segmented and identified using deep networks, while anatomic consistency is enforced using statistical priors. In this strategy, the transitional vertebrae identification is handled by encoding their configurations in the proposed graphical model that aggregates local deep network predictions into an anatomically consistent final result. Furthermore, our method can detect and report inconsistent spine regions that do not satisfy the anatomic consistency priors. The code and model are made available for research purposes [1].

The tasks of vertebrae localization, segmentation and identification are intrinsically inter-dependent and a common issue comes from inconsistencies that can propagate between them. For instance, neighboring vertebrae share similar shapes which makes their identification uncertain. Moreover, pathological spines can present abnormal shapes or the number of transitional vertebrae can be different among patients. These transitional

[1] https://gitlab.inria.fr/spine/vertebrae_segmentation

vertebrae, i.e. , the absence of T12 or occurrence of T13 or L6, are common and are reported to affect between 15% and 35% of the general population [21, 113, 65]. However, they only impact one vertebrae among the 24 in a spine and their occurrence percentages in standard vertebrae datasets appear to be therefore much lower, down to a few percentages in practice. As a consequence, the current state-of-the-art methods, which are mostly based on deep networks trained over standard datasets, tend to experience important performance drops in the presence of transitional vertebrae.

Besides augmenting datasets, an alternative solution is to exploit prior knowledge on the full structure of the spine, as initially introduced in non-learning based methods. In this work, we investigate the combination of such a strategy with a deep learning approach. Precisely we propose to iteratively cycle between the localization, segmentation and identification tasks, which uses deep networks, while enforcing anatomical consistency with statistical priors. On the one hand, the priors are used to localize vertebrae. They leverage learned statistics of vertebrae volume and inter-vertebral distances which exhibit more robustness to pathological cases than shape or appearance models. On the other hand, we proposed to encode the admissible configurations in a graphical model, in order to handle transitional vertebrae in the identification.

5.2 ANATOMIC CONSISTENT CYCLE

Our method starts by segmenting a spine binary segmentation, so called *spine mask* (Chapter 3), which allows to locate individual vertebrae (Sec. 5.2.1). From the obtained locations, individual vertebrae segmentation masks are estimated with an individual vertebrae segmentor (Sec. 5.2.2) and refined with an iterative location-segmentation refinement scheme (Sec. 5.2.2) . Vertebrae are further identified using the locations and segmentation masks (Chapter 4). The obtained identifications allow, in turn, to enforce finer anatomic consistency constraints (Sec. 5.2.1) and to detect possible new candidate locations. In the latter case, the new vertebrae go through the segmentation and identification steps described before. The process ends when the proposed consistency criteria are satisfied. It also stops when the set of detected locations, segmentation masks and identifications does not change over the cycle. The remaining inconsistencies, such as a failure in the location-segmentation refinement or an inconsistency in the anatomic constraints, are reported in the result. The overview of the method is illustrated in Fig. 5.1.

5.2.1 ANATOMIC CONSISTENCY CONSTRAINTS FOR VERTEBRAE LOCALIZATION

Vertebrae are naturally ordered in the spine and their relative locations and consecutive sizes are heavily correlated. With the objective to exploit such prior information we compute statistics over inter-vertebral distances and individual vertebrae volumes. Given these priors we enforce then two consistency criteria. First, distances between all detected locations should follow the inter-vertebral distances statistics. Second, the *spine mask* should be similar to the union of individual vertebrae masks. The segmentation of a *spine mask*

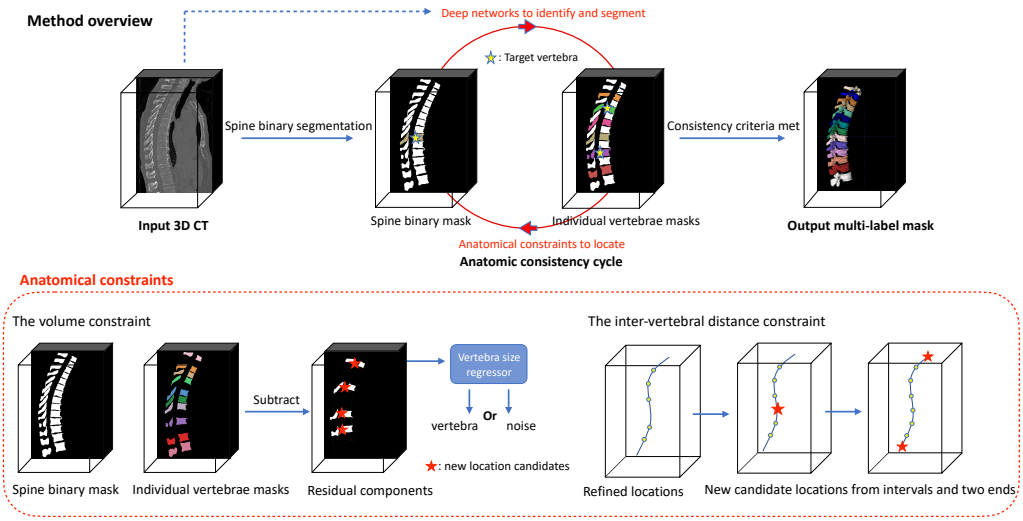


Figure 5.1: The method overview. Given a 3D CT as input, the spine is segmented as a reference to locate the individual vertebrae. The anatomical constraints are leveraged with deep networks for localization and segmentation. Once the location is stable, its identification is obtained given its segmentation. The set of location, segmentation and identification is cycled through until the consistency criteria are met.

is, in general, a straightforward task which in turn allows to identify the areas where individual vertebrae should lie. In practice, at each cycle the difference between the *spine mask* and all individual vertebrae masks is computed to obtain residual connected components. (The residual connected component is the whole *spine mask* at the first cycle.) Vertebrae volume statistics are then considered to decide whether a residual is a vertebra or just noise.

The vertebrae volume constraint Given a residual connected component, the idea is to check its volume by considering the volumes of the neighboring vertebrae. To this purpose a regressor is trained over consecutive vertebrae to predict the volume of the previous or next vertebrae along the spine.

If the residual volume size is larger than a fraction of the predicted size (50% in our experiments), it is considered as a vertebra and its location at its mass center is added to the vertebrae list. The residual is otherwise discarded as noise. When the vertebra identification is available, one can condition the regressors by spine levels, which improves the accuracy of the prediction. Although the described volume constraint allows to locate vertebrae, it should be noticed that if the *spine mask* presents failures, i.e. on abnormal vertebrae, locations can still be missed.

The inter-vertebral distance constraint. This constraint builds on the the fact that distances between locations are well structured. To this aim, we use two statistical models of the distance between vertebrae. The first is a Gaussian distribution for each anatomic group capturing the mean and variance of the distances between consecutive vertebrae. It

is used as a prior to detect abnormally large distances. In addition, linear regressors are used to predict the inter-vertebral distance from the neighbouring ones. In practice they predict the inter-vertebral distance from: i) the previous vertebrae distance, ii) the next one and iii) using both sides. These regressors are also conditioned on each anatomic group. When a larger inter-vertebral distance is detected, new candidate locations are added in between. The number of new candidates is adapted depending on the current distance and the predicted one.

We also leverage identification to check if the spine extremes are complete. If C1 (or L5) is not yet found, a location is predicted up (or down) using the two most top (down) locations. If it is inside the image field of view, it is added.

5.2.2 INDIVIDUAL VERTEBRA SEGMENTATION

Individual vertebra segmentor For the individual vertebrae segmentation, we use the same strategy described in Chap. 3 taking as input a crop of the input CT image and one 3D vertebra location in a two-channel 3D image [94]. For the location we use a 3D Gaussian map centered at the 3D location with a $\sigma = 20$. The location is included to segment the target single vertebra while ignoring the neighboring vertebrae. To be noticed that, because the image is centered and cropped, the second channel is always the same. The output is a one-channel probability map which is then binarized with a 0.5 threshold to obtain the final individual vertebra binary mask. Different from the spine binary segmentation (Chap. 3), we use an input cube size of $128 \times 128 \times 128$ being able to cover any full shaped vertebra in any orientation. We train the segmentation network to minimize Eq. 3.1 with $\lambda = 20$.

Although the input is in PIR orientation, some vertebrae, specially the cervical, have important rotations around the Z-Axis. In addition, the input 3D locations might not be accurate. To make the network robust to these cases we augment the data with locations translated with $t \in [-10, +10]^3 mm$ as well as with rotations in $[min(-50, \theta - 50), max(50, \theta + 50)]$ degrees around the Z-Axis, where θ is the vertebra rotation wrt the axial plane. The angle θ is computed using the current and next vertebrae locations. Each sample is augmented 162 times.

Iterative location-segmentation refinement From the previous paragraph, an individual vertebra can be segmented given a location. Further, the individual vertebra location can be approximated from its shape by computing the center of mass of the segmentation mask [70]. We quantify the accuracy of the heuristic that is used when computing the location of a vertebra as the center of mass of the segmentation mask. For each GT vertebra mask in the training set, we compute its gravity center, We then compute the Euclidean distance between each gravity center and its GT location and obtain a mean distance error of 0.67 ± 0.59 mm. We consider this sub-pixel precision accurate enough to validate the approach. This heuristic produces a fairly precise estimate of the vertebra location, provided the quality of the segmentation mask is qualified.

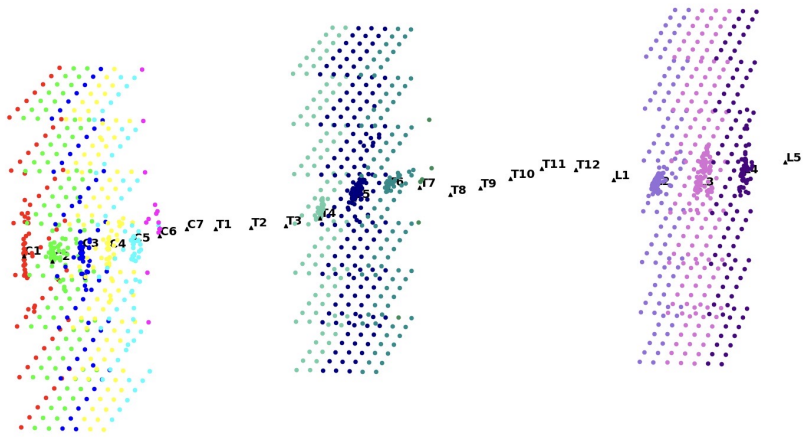


Figure 5.2: Bassin of attraction of the vertebra location. We show the initial locations sampled around C3, T5 and L3, their converging waypoints and converged locations. (Best viewed in color)

To better understand the entanglement of the location-segmentation and the *bassin of attraction* of each vertebra, we perform an experiment where we sample the initial location on a 60 mm grid around a GT location with a stride of 10 mm. In Figure 5.2 we illustrate the visuals of the sampling and convergence trajectory around three vertebrae C3, T5 and L3. Each vertebra is associated with one color and we color-code each initial location with the color of the vertebra to which the refinement converged. For instance, red represents C1 and the red dots mark the initial locations that converged to the C1 vertebra. We note how the initial locations color coding is structured in space, neighboring initial locations converged either to the same vertebra, or the neighboring one. It is also worth noting that the size of the *bassin of attration* differs for cervical, thoracic and lumbar vertebrae. Its size is clearly related to the size of the vertebrae. The thoracic and lumbar vertebrae have a larger size than the cervical ones.

Thus, from a given location, the individual vertebra segmentor is able to segment a single vertebra mask and separate it from the neighboring vertebrae. Then, one can compute the center of mass of the predicted individual segmentation and get a new location. This cycle can be then iterated until the convergence of the location and the individual segmentation mask is reached. A refined pair of vertebra location and segmentation is obtained. We study the convergence of the refinement process: we initialize a location with a perturbation from the GT location, run the iterative refinement and observe the results. We initialize the location with different settings: GT, GT+5 mm, GT+10 mm and GT+15 mm. For GT, the converged locations end up with 1.19 mm mean distance error and converged masks achieve 91.43% DSC. For the perturbation of 5 mm, 10 mm and 15 mm, the converged location errors are 1.22 mm, 1.21 mm, $1,21$ mm and the converged masks obtain a Dice score with 91.37%, 91.67%, 91.62% respectively. In Figure 5.3 we show the case of GT+15 mm how the distances between the locations of consecutive iterations get smaller (left), and how the overlap of consecutive segmentation masks becomes stable

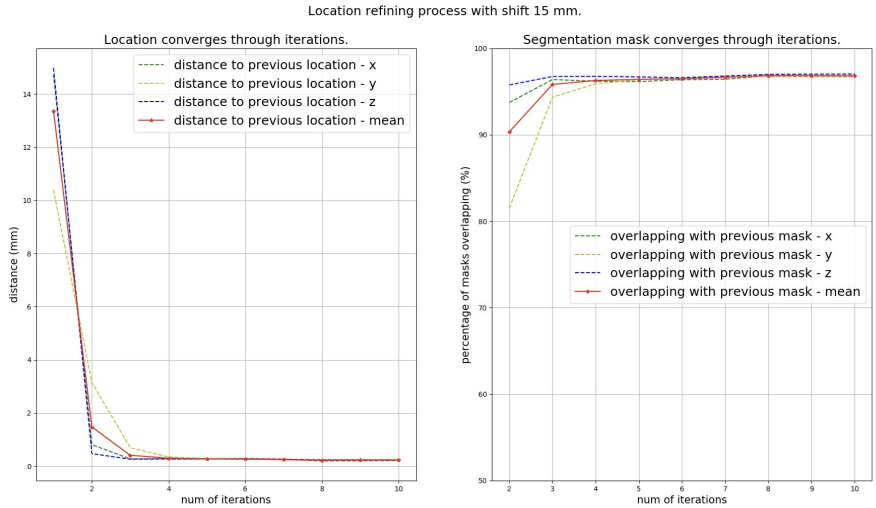


Figure 5.3: Vertebrae location-segmentation refinement. X axis represents the number of iterations. Left: locations convergence. Y axis represents the Euclidean distance to the previous locations. Right: segmentation convergence. Y represents the percentage of the overlapping with the previous segmentation mask.

(right).

Since the segmentation masks of the same vertebra always converge to the same location, this location-segmentation refinement scheme can also used to detect duplicated locations. Extra detections can just be removed by keeping one pair of location and segmentation of the vertebra.

5.3 VERTEBRAE VOLUME AND INTER-VERTEBRAL DISTANCE STATISTICS

For locating the vertebrae, we use the statistics of vertebrae volume and inter-vertebral distances. They follow the nature of the human spine and the consecutive ones are highly correlated. The vertebra volume represents the shape of the individual vertebra while it is more robust to morphological deviations than the surface model or image appearance priors. The inter-vertebral distance quantifies the spine structure and implicitly reflects the discs statistics. A narrow distance indicates it may be from the cervical level or the patient is younger, and a large distance implies the lumbar area or elder patients. Abnormally large distance raises the alert of possible missing vertebrae detection in our work.

Vertebrae volume regressor Vertebra volume, which we define as the number of voxels of the vertebra in a volumetric-grid representation of resolution $1 \times 1 \times 1\ mm^3$, vary across the cervical, thoracic and lumbar while the consecutive vertebrae volumes are heavily related as shown in Fig 5.4. The statistics can be employed to learn a linear regressor that predicts a vertebra volume given the neighboring one. To be specific, for each anatomic

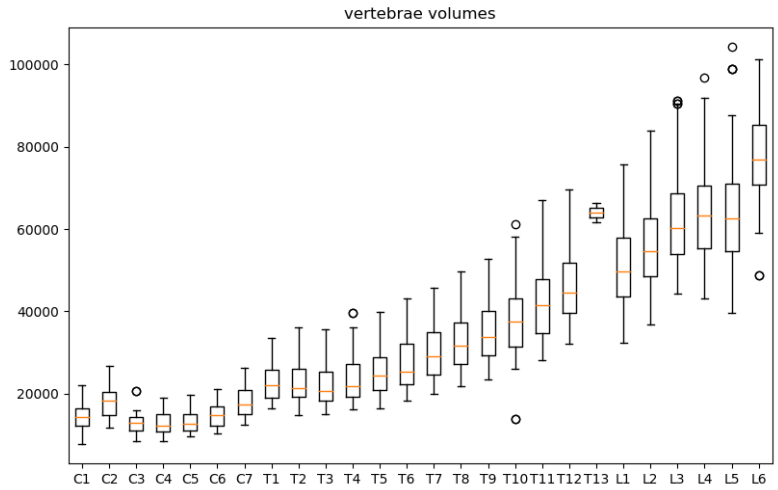


Figure 5.4: The boxplot of the vertebrae volumes from C1 to L6. An increment is observed from the cervical to lumbar vertebrae volumes. Inside each spine level, the distribution can be estimated as linear.

	a	c_1	b	c_2
cervical	1.03	1471	0.92	497
thoracic	1.03	1354	0.94	-140
lumbar	1.05	981	0.94	-269

Table 5.1: The learned coefficients of vertebrae volume regressors

group (cervical, thoracic and lumbar), two linear regressors are trained to predict the vertebra volume given the previous and the next vertebra respectively:

$$S_i = a * S_{i-1} + c_1 \quad and \quad S_i = b * S_{i+1} + c_2 \tag{5.1}$$

where the S represents the quantity the vertebra volume. a, b, c are the coefficients of the linear regressor, the learned values are shown in Tab 5.1.

Inter-vertebral distance regressor The vertebrae locations are consistent and well structured. On the one hand, the inter-vertebral distances are restricted in a range and strictly follow a distribution among the human. On the other hand, one distance is highly correlated to its neighboring ones. Based on this prior, we learn the inter-vertebral distances in two forms. One is the Gaussian distribution of each anatomic group, the other is linear regressors predicting the inter-vertebral distance given either the previous, next or both-side distances for each anatomic group.

The Gaussian distribution of the inter-vertebral distances is represented as

$$d(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{5.2}$$

	μ (mm)	σ (mm)
cervical	16.77	2.18
thoracic	23.32	3.55
lumbar	32.68	2.84

Table 5.2: The learned coefficients of Gaussian distributed inter-vertebral distances.

	m_1	n_1	k_1	m_2	k_2	n_2	k_3
cervical	0.55	0.45	-0.08	0.92	2.40	0.98	-0.13
thoracic	0.57	0.44	-0.24	0.93	2.29	0.97	-0.07
lumbar	0.56	0.46	-0.73	0.95	1.96	0.96	0.23

Table 5.3: The learned coefficients of the inter-vertebral distance regressors.

The mean μ and standard deviation σ of each anatomic group are shown in Tab 5.2. Increasing distances between consecutive vertebrae are observed from cervical to lumbar level. The distance $d(x)$ is considered as in the reasonable range when

$$\mu - 3\sigma < d(x) < \mu + 3\sigma \tag{5.3}$$

The learned Gaussian distribution is less flexible as some patients may have larger inter-vertebral distances and kids' are much smaller. We then learn to predict the inter-vertebral distance given the neighboring ones. Three types of regressors are learned: using the both-side neighboring distances

$$G_i = m_1 * G_{i-1} + n_1 * G_{i+1} + K_1, \tag{5.4}$$

using the previous and next distance

$$G_i = m_2 * G_{i-1} + K_2 \quad and \quad G_i = n_2 * G_{i+1} + k_3. \tag{5.5}$$

The latter two are intended for the most top and bottom vertebrae who do not have the distances from both sides. The learned coefficients are presented in Tab. 5.3.

To evaluate the regressors of the inter-vertebral distance, we compute the Mean Square Error (MSE) between the ground truth relative distances and the distances predicted by the regressors. We show the mean μ and standard deviation σ of MSE in Tab 5.4. An inter-vertebral distance is considered as *normal* when

$$\mu_{mre} - 3 * \sigma_{mre} < MRE < \mu_{mre} + 3 * \sigma_{mre} \tag{5.6}$$

5.4 VERSE CHALLENGE BENCHMARK

We quantitatively evaluate our method on the VerSe20 Challenge public and hidden test-sets with 200 CT subjects in total and adopted the metrics of the challenge evaluation protocol [102]: The *Dice coefficient (DSC)* and the *Hausdorff Distance Error (HD)* evaluate

	both sides		previous		next	
	μ_{mre}	σ_{mre}	μ_{mre}	σ_{mre}	μ_{mre}	σ_{mre}
cervical	9.13	2.86	10.20	2.05	12.13	3.36
thoracic	2.42	1.43	3.96	1.56	4.17	0.88
lumbar	2.04	0.93	4.45	1.55	5.16	1.71

Table 5.4: The mean and standard deviation of MSE of the inter-vertebral distance regressors.

	public testset				hidden testset			
	labelling results		segmentation results		labelling results		segmentation results	
	ID rate(%)	MLD(mm)	DSC(%)	HD(mm)	ID rate(%)	MLD(mm)	DSC(%)	HD(mm)
Zhang A.	94.93	2.99	88.82	7.62	96.22	2.59	89.36	7.92
Payer C.	95.06	2.90	91.65	**5.80**	92.82	2.91	89.71	**6.06**
Chen D.	95.61	**1.98**	91.72	6.14	**96.58**	**1.38**	**91.23**	7.15
ours	**96.59**	2.22	**92.50**	7.09	95.24	2.43	91.10	6.72
ours(a)	96.02	2.22	92.06	6.82	94.93	2.34	90.85	6.53
ours(b)	93.61	3.10	89.76	7.80	89.41	3.81	85.45	8.39

Table 5.5: Quantitative comparison of our method with the top ranked methods on the VerSe20 challenge[102] as well as with alternative (a) in which the anatomic consistency constraints are not conditioned on the identification and alternative (b) where the anatomic consistency priors are not used for detection.

the correctly identified segmentation performance in terms of voxel and surface similarity respectively. The *Identification rate (ID rate)* and the *Mean Localization Distance (MLD)* evaluate the accuracy of the labelling task, where MLD measures the Euclidean distance of the predicted location to the GT location and 20 mm is the criterion for a valid identification. According to the challenge metrics, when a vertebra is missed, MLD and HD are not computed and accounted in the final scores averaging. Thus there is a trade off between ID rate and MLD (same for DSC and HD). While more positive detections lead to a higher ID rate, a worse MLD can be obtained as the detection errors are taken into account.

Table 5.5 shows the results of the proposed method and the top scoring methods in the benchmark. Our method is on pair with the best performing method in the leaderboard [24]. It is worth noting that the method by Chen D. et al. [24] did not win the challenge, as one important criterion in the challenge was the performance in handling the transitional vertebrae - 6 cases with T13 (2/2/2 in Train/Public/Hidden), 47 cases with L6 (15/15/17) and 8 cases with absent T12(3/4/1). Following the challenge guidelines we computed the Dice score only on scans with transitional vertebrae and the obtained results are presented in Table 5.6. Our method consistently outperforms all existing methods.

Figure 5.6 shows qualitative results of the method. Severely fractured vertebrae (T8, T9 and L1) are well segmented and identified in Fig. 5.6-(a). Metal-inserted vertebrae are successfully handled in Fig. 5.6-(e,f). In Fig. 5.6-(b), the location-segmentation refinement [70] fails due to the occurrence of metal and T4 is not segmented. Thanks to the anatomic priors, T4, T5 and T6 can still be properly located and identified. The cross next to the label indicates the detected inconsistency. The transitional vertebrae such as

	public testset	hidden testset
ours	**91.04**	**89.70**
Payer C.	85.96	89.59
Zhang A.	87.15	87.35
Chen D.	84.21	87.01

Table 5.6: Methods evaluation (Dice score %) on transitional vertebrae. [102]

	DSC(%)
ours	**90.84**
Chen D.	86.44
Payer C.	84.11
Zhang A.	85.42

Table 5.7: Generalization performance on VerSe19 hidden testset. [102]

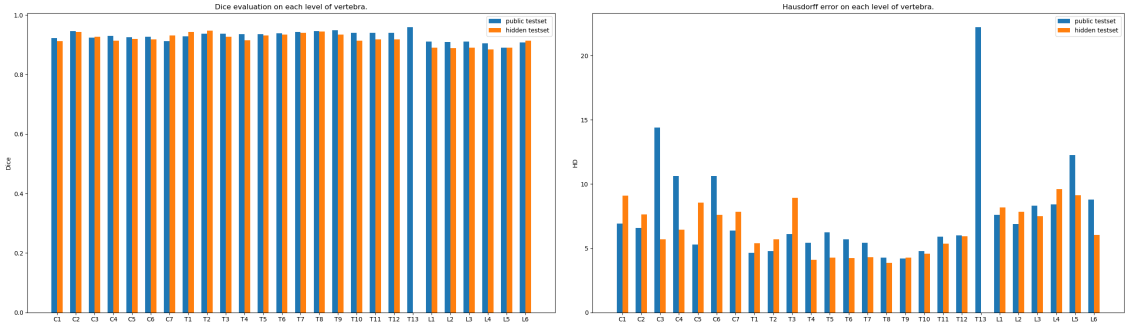


Figure 5.5: Evaluation on individual vertebra. Left: Dice score (%); Right: Hausdorff distance (mm).

L6 and T13 are detected in Fig 5.6-(c,d).

The dataset is class imbalanced that thoracic and lumbar vertebrae are present more than the cervical vertebrae. To evaluate the performance on each level of vertebra, we compute the Dice score and Hausdorff distance error on individual vertebra. As shown in Figure 5.5, the achieved accuracy is quite balanced for all the vertebrae except the transitional vertebra T13, where it fails to be detected in hidden testset.

To evaluate the generalization of our method, we perform the experiment of the Verse20 challenge [102] where the approach developed and trained on the VerSe20 training set is applied to the VerSe19 hidden testset. Table 5.7 shows that our method outperforms all approaches and generalizes well to VerSe19.

Ablation study To highlight the benefits of the different parts of the proposed method, we perform two experiments. In the first one (*ours(a)*) in Tab. 5.5 we do not use the identification to condition the vertebrae statistics. Instead of learning specialized regressors for each spine level, we use a fixed threshold to determine if a residual mask is a potential true vertebra (half of the smallest vertebra volume size in the training set is selected: $7820/2$). To decide if there is a missing vertebra given a distance between 2, we use a fixed threshold of 50 mm [94]. The overall scores show that the improvement when using conditional statistics is minor but not negligible: without conditioning 8 vertebrae are missed in the public testset and 6 in the hidden testset while with the identification information no vertebra is missed in public testset and only 4 in the hidden testset. In the second experiment (*ours(b)*) in Tab. 5.5, the anatomic consistency priors (Sec. 5.2.1)

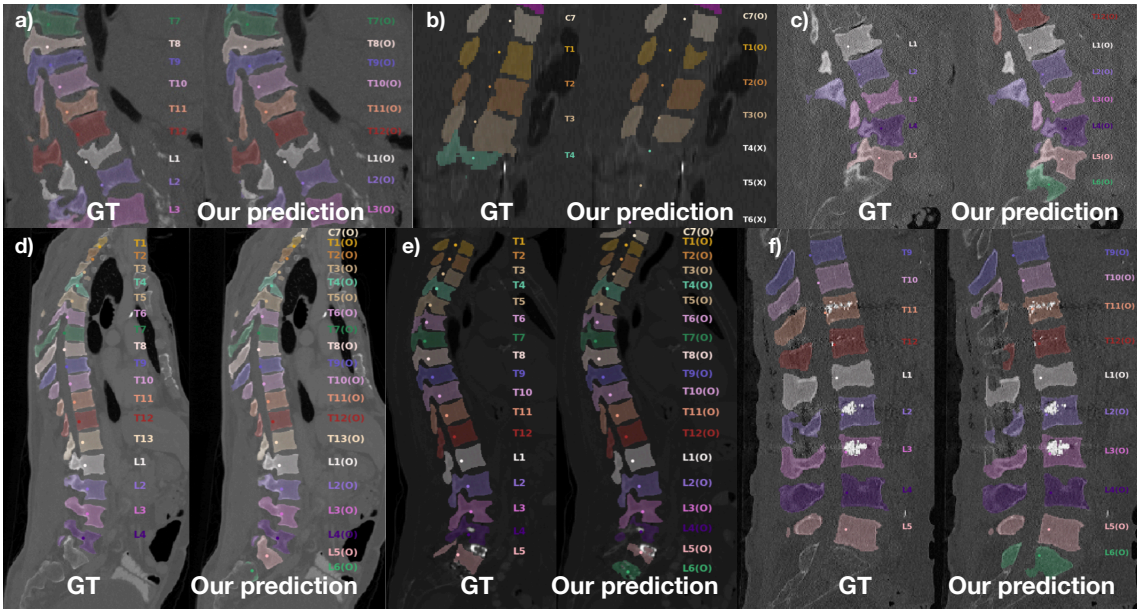


Figure 5.6: Qualitative results on fractured (a), metal-inserted (b,e,f) and transitional (c,d) vertebrae.

are not used. Without the constraints on vertebrae volume and the relative locations, the performance significantly drops.

5.5 Failure Cases and Results Visualization

We inspect the results that obtain Dice scores lower than 90% and observed that failures of the method primarily come from wrong vertebrae identification, typically the one label shift. Figure 5.7 shows the confusion matrix of the final identifications. It contains 26 classes: 24 vertebrae from C1 to L5 and the transitional vertebrae T13 and L6. Most confusions arise in the thoracic-lumbar transition.

Two typical scenarios are observed.

First, all the vertebrae are identified with one label shifted as shown in Fig. 5.8. The confusions are around thoracolumbar area. To investigate if the identification error is from the individual vertebra prediction or the graph optimization, we inspect the probabilities of the individual prediction. We found that the individual vertebra classification network holds high confidence in its outputs. Take the case c) in Fig. 5.8 as an example, the labels from top down predicted by the network are [17, 18, 19, 20, 21, 22, 23, 24, 25][2] with probabilities [1.0, 1.0, 1.0, 0.99, 1.0, 1.0, 1.0, 0.99, 1.0], while the GT labels are [18, 19, 20, 21, 22, 23, 24, 25]. As we know, one of the challenges of vertebrae identification is the high similarity of the neighboring vertebrae shapes that would confuse the deep neural network. On the contrary, the network is much confident about its prediction without any

[2]The numerical labels 1-25 correspond to the vertebral levels C1-L6

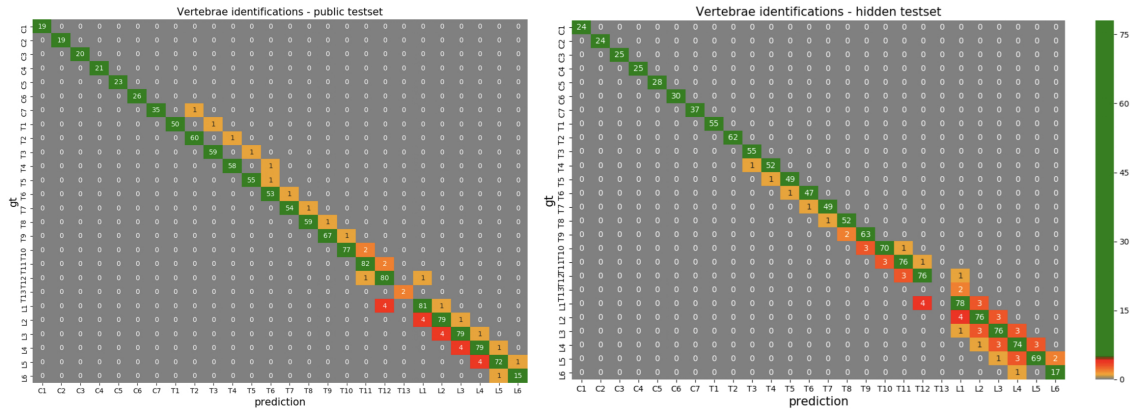


Figure 5.7: Final results confusion matrices. Left: public testset; Right: hidden testset.

uncertainty between options. It implies that either the network is not capable enough to extract the implicit features of each vertebra, or the network is not able to classify and distinguish between them. The drawback points to the network architecture design, where we use a baseline backbone architecture *3D vgg16* (Chap. 4). Further improvement can be made towards to the optimization of the individual vertebra classification network.

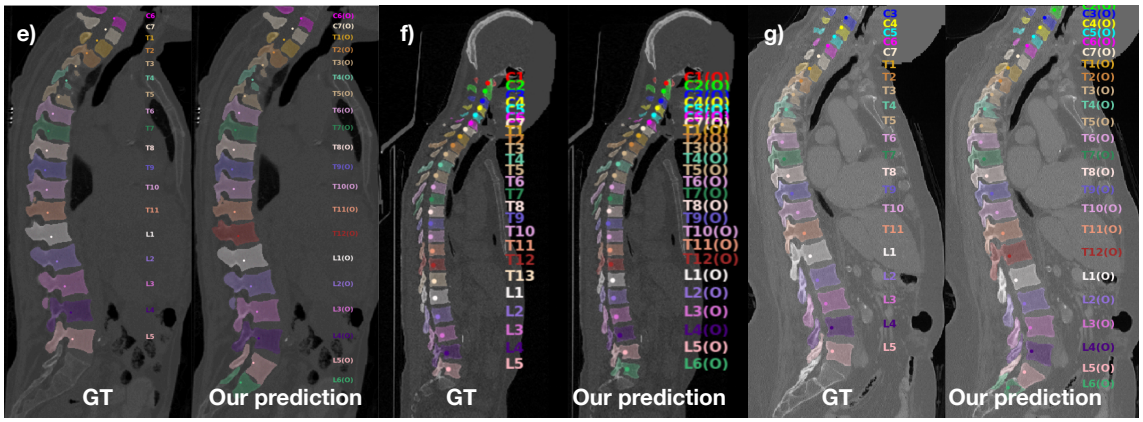


Figure 5.9: Failure cases on transitional vertebrae.

Second, the identification fails on challenging transitional vertebra and the consequent labels are shifted as shown in Fig. 5.9. The graph explicitly models the transitional vertebrae. Nevertheless, it is still a challenging task. One possible reason for the failure is because of the imbalance of the presence of the transitional vertebrae in the training and test sets. Transitional vertebrae are rare in the dataset. Moreover they are unproportionally distributed in the training and test sets to challenge the algorithm. In our method, the cost of bridging the abnormal transitions is higher and weighted, where the weights for the three transitions (the presence of T13, L6 and the absence of T12) are turned with the training set. It does not generalize well to the test sets. The improvement can be explored in learning the weights for the transitional vertebrae and the other individual vertebrae.

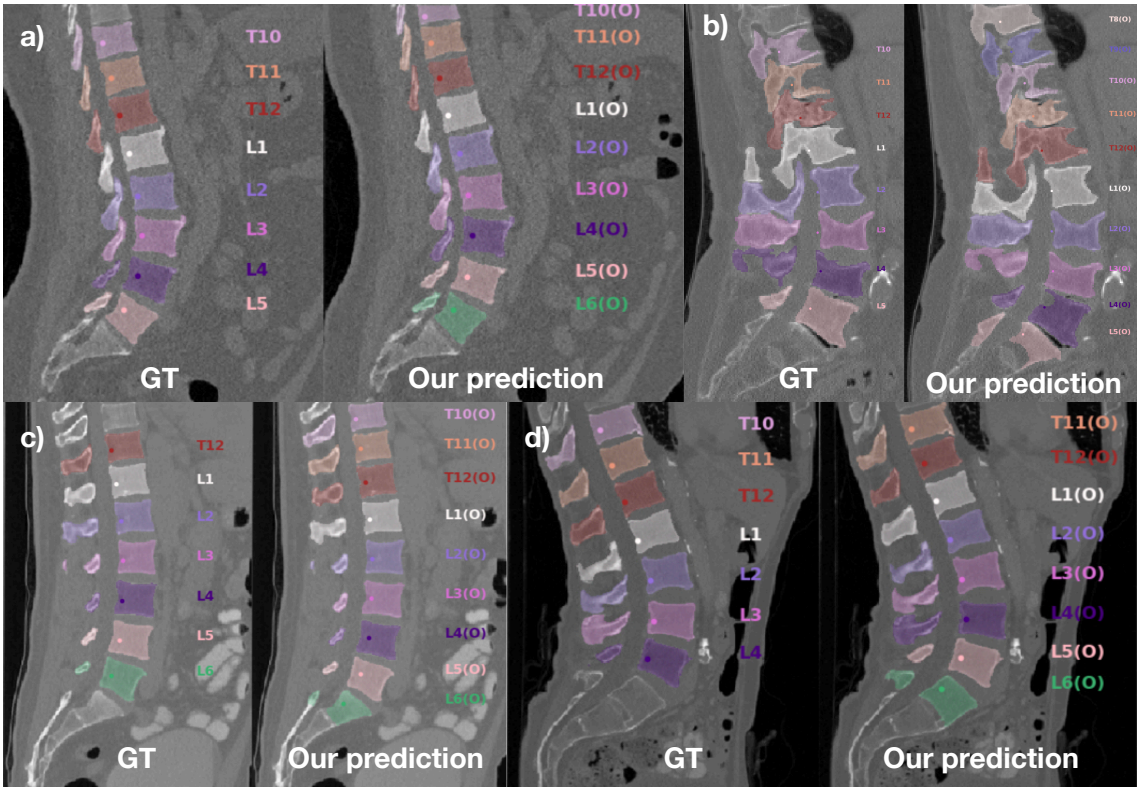


Figure 5.8: Failure cases with one label shifted.

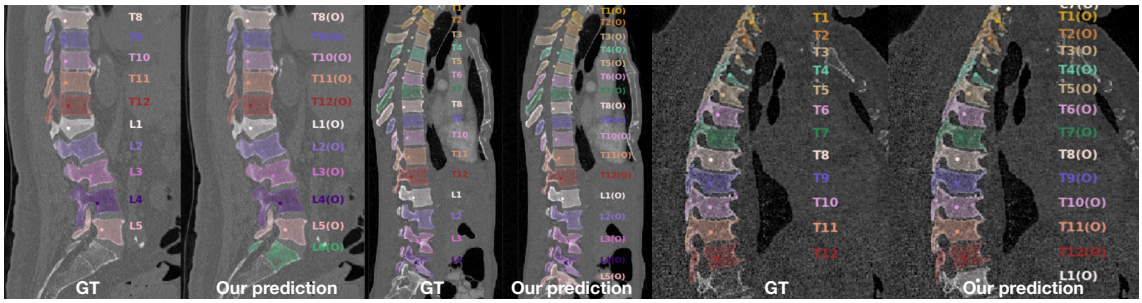


Figure 5.10: Results visualization of segmentation on fractured vertebrae and poor quality scans.

We further provide some visualizations of the results both from public and hidden test-sets. Figure 5.11 shows the general segmentation performance on cervical vertebrae. Figure 5.12 shows the full spine where our algorithm segments until the bottom of the spine including L6. Fractured vertebrae and poor-quality CT scans can be found in Figure 5.10, our algorithm is able to deal with compressed vertebrae, distorted shaped vertebrae as well as the scan of poor quality. In general, besides the area that labeled in the ground truth mask, our prediction also handles the corner cases such as the vertebrae at the image border. Moreover, a report of the inconsistent spine regions is provided along with the segmentation and labelling result.

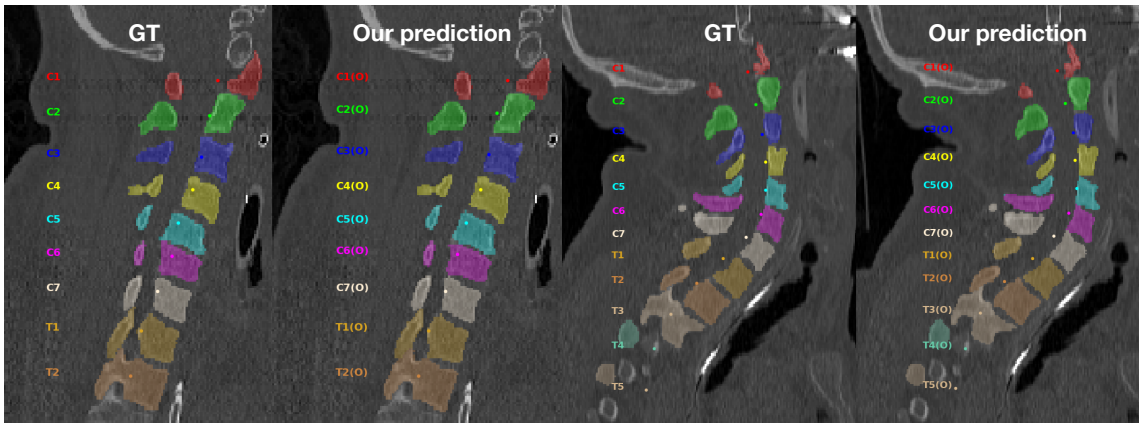


Figure 5.11: Results visualization of segmentation in the field of view of cervical vertebrae.

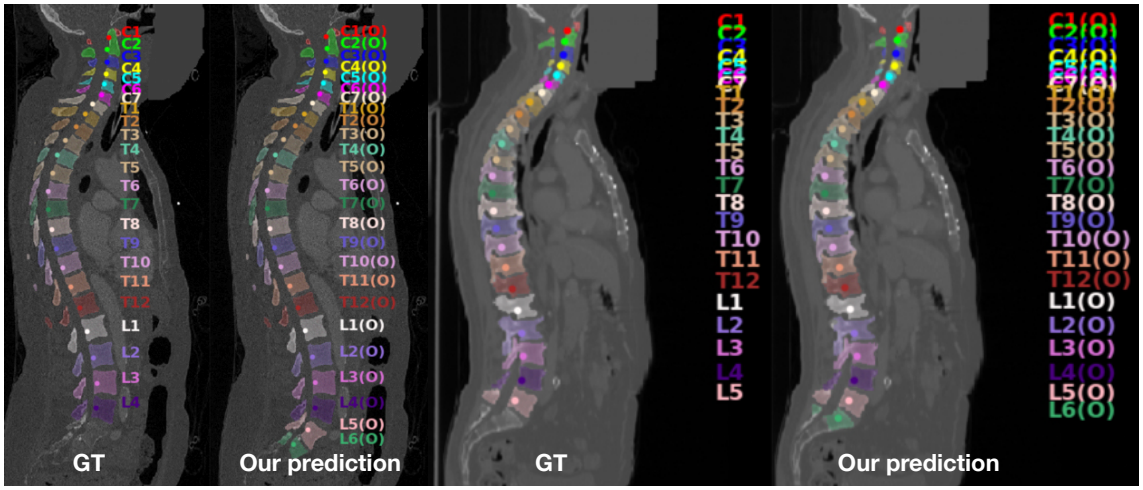


Figure 5.12: Results visualization of segmentation in the field of view of full spine.

5.6 CONCLUSION

In this chapter, we present a new approach to localize, segment and identify vertebrae in CT images. An anatomic consistency cycle is proposed that aggregates task-specific deep networks while enforcing statistical priors. As seen with the generalization on the VerSe19 dataset, it improves robustness in the results. The evaluation on the standard VerSe20 challenge benchmark demonstrates the interest of the proposed strategy, in particular with transitional vertebrae which are present in a meaningful part of the population. The extension of the proposed anatomic consistency cycle to other anatomic structures with similar specific cases is currently under investigation.

CHAPTER 6

GENERALIZATION AND LIMITATION

6.1 GENERALIZATION

Previous chapters demonstrate our proposed method of segmenting and identifying vertebrae from CT images and its promising results on VerSe'20 dataset [102]. Moreover, it achieves the state of art performance on generalizing to the unseen VerSe'19 testset. To further access the generalization ability of the method, we evaluate it on several publicly available datasets and compare the performance with the state of the art methods. These datasets are from the *SpineWeb*[1] and consisted of 3D CT images of the spine.

xVertSeg dataset The xVertSeg dataset [49] , released as part of the xVertSeg challenge in MICCAI 2016, consists of 15 train CT volumes with ground truth segmentation of the lumbar vertebrae (into five classes, L1-L5) and 10 test CT volumes. The ground truth segmentation of the test set is not publicly available. The dataset includes non-fractured vertebrae and vertebrae with fractures of different morphological grades and cases. The scans were reconstructed to in-plane resolutions of 0.29 mm to 0.80 mm and slice thickness of 1.0 mm to 1.9 mm. The scans are with varying field of views, while mostly observing the lumbar area.

To evaluate our method, we use the scans 1 to 5 for evaluation and the remaining 10 scans for training, which is the same setting introduced in [70, 29]. Therefore, we can compare our method to theirs. The quantitative results are shown in Table 6.1. We obtain mean Dice score 94.65% which is comparable to the approach of Lessmann et al. [70] and superior than that of Chuang et al. [29]. Concerning the average absolute symmetric surface distance (ASSD), our result obtains a larger distance error 0.6 ± 0.1 mm than Lessmann et al. [70]. Figure 6.1 shows the qualitative results, we can see the fractured vertebrae are well segmented and identified and the thoracic vertebrae that in the field of view but not annotated in the ground truth are also segmented, as our method works for arbitrary field-of-view CT scans.

[1] `http://spineweb.digitalimaginggroup.ca`

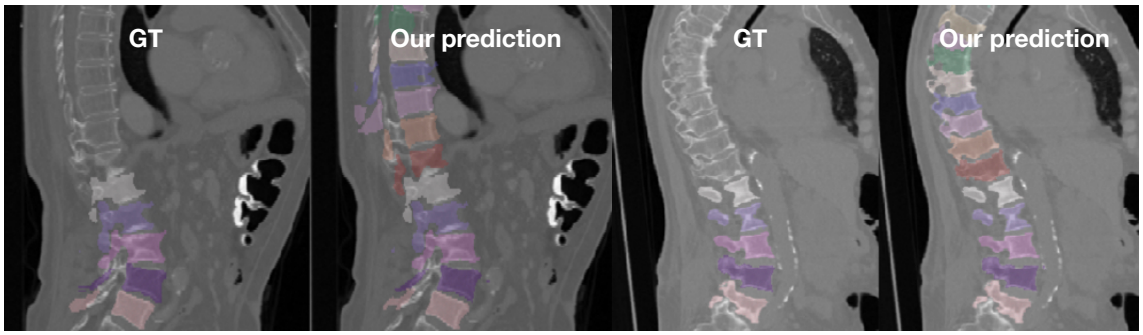


Figure 6.1: Results visualization of xVertSeg dataset.

	Dice(%)	ID rate(%)	ASSD(mm)	HD(mm)
ours	94.6±0.9	100	0.6±0.1	13.06±6.71
Lessmann et al.[70]	94.6±2.2	100	0.3±0.2	-
Janssens et al.[53]	95.7±0.8	100	0.4±0.1	4.32±2.60
Sekuboyina et al.[104]	94.3±2.8	-	-	-
Chuang et al.[29]	88.5	100	-	-
Cheng et al.[27]	87.7±3.5	100	-	-

Table 6.1: Quantitative results of xVertSeg dataset.

There are other methods that are evaluated using the same dataset, but with different evaluation/training protocols. Janssens et al. [53] conducted a leave-three-out cross-validation study to evaluate the performance of their proposed method. More specifically, each time they randomly took 3 out of 15 CT data as test data and the remaining 12 CT data as the training data. The process was repeated for 5 folds. Sekuboyina et al. [104] used the 15 training set to train their model and evaluate the performance on test data. As the ground truth annotation of the test data is not released, they opted for an in-house ground truth generation. The near-perfect segmentation from their approach was given to two clinical experts (Rater-1 and Rater-2) for correction. Rater-1 was tasked to correct the entire volume, while Rater-2 was tasked to pick a random subset of sagittal, coronal and axial slices from a volume and segment them entirely. Using the ground truth generated from Rater-1, they obtained $94.3\% \pm 2.8\%$ Dice accuracy. $92.0\% \pm 2.3\%$ Dice score was obtained when using Rater-2 ground truth segmentation. Cheng et al. [27] divided the 15 training data into two parts: 10 images for training and 5 for testing. The split information was not explicitly provided.

Lumbar vertebra segmentation CT image database The LumbarSeg dataset consists of 10 scans of healthy subjects and associated annotation for lumbar vertebrae. The slice resolution is between 0.28 mm to 0.79 mm and the slice thickness is between 0.72 mm to 1.53 mm. We follow the evaluation protocol of Lessmann et al. [70] that use it for an external evaluation of our supervised method. Scans from this dataset were only used for evaluation and were not part of the training set.

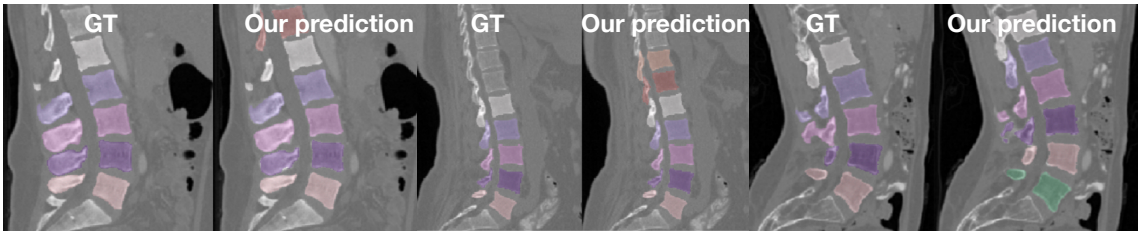


Figure 6.2: Results visualization of LumbarSeg dataset.

	Dice(%)	ID rate(%)	ASSD(mm)	HD(mm)
ours	95.3±0.3	90	0.3±0.1	7.4±1.5
Lessmann et al.[70]	96.5±0.8	100	0.2±0.0	-
Korez et al.[67]	95.3±1.4	-	0.3±0.1	-
Chu et al.[28]	91.0±7.0	-	0.9±0.3	7.3±2.2
Ibragimov et al.[50]	93.6±1.1	-	0.8±0.1	-

Table 6.2: Quantitative results of LumbarSeg dataset.

AS shown in Table 6.2, by applying our model on the unseen dataset we obtain Dice score (%) of 95.3 ± 0.3 which is consist with the performance on VerSe20 dataset. The accuracy is 1.2% lower than Lessmann et al. [70]. For the identification accuracy, we obtained 45 correctness out of 50 vertebrae. The "one label shift" issue is found in one case, shown in Figure 6.2 right. L5 is recognized as L6 causing the rest vertebrae above it are with one label offset. This is the main failure of our method and it was also presented in the VerSe20 dataset evaluation. It is due to the individual vertebra classification network that the similarity of neighboring vertebrae is challenging the backbone architecture. Korez et al. [67] achieved on-par performance when using a different evaluation protocol, where they split the 10 CT subjects into train and test sets. Chu et al. [28] evaluated their method with a leave-one-out cross validation. Only vertebral bodies were segmented and evaluated.

CSI2014 segmentation dataset The CSI2014Seg dataset, released as part of the spine segmentation challenge in MICCAI 2014, consists of 15 CT scans of healthy subjects. Full thoracic and lumbar vertebrae (17 vertebrae in total) are observed and their corresponding segmentations are provided. The scans were reconstructed to in-plane resolutions of 0.31 mm to 0.36 mm and slice thickness of 0.7 mm to 1.0 mm. The dataset is split into training set of 10 scans and test set of 5 scans. We evaluate our method following the challenge protocol where 10 CTs were used for training and 5 for testing.

Our method obtained Dice score (%) of 93.4 ± 0.8 , a bit lower than Lessmann et al. [71, 70] and Cheng et al. [27] who used the same train/test split. Results visualization can be found in Figure 6.3. For other approaches, they use only the 10 training data. Hammernik et al. [44] and Korez et al. [67, 68] performed a leave-one-out 10-fold cross validation and reported the average accuracy over 10 experiments. Kolavrik et al. [64]

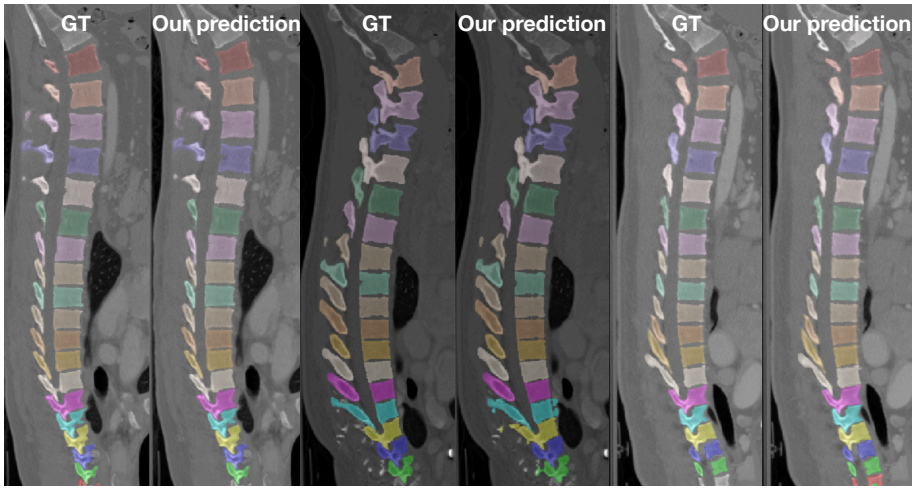


Figure 6.3: Results visualization of CSI2014Seg dataset.

	Dice(%)	ASSD(mm)	HD(mm)
ours	93.4±0.8	0.7±0.1	6.6±1.2
Kolavrik et al.[64]	97.1±0.0	0.3±0.1	-
Lessmann et al.[1][70]	96.3±1.3	0.1±0.1	-
Lessmann et al.[2][71]	94.8±1.6	0.3±0.1	-
Cheng et al.[27]	95.3±1.4	-	4.0±2.1
Korez et al.[1][67]	94.6±2.0	0.3±0.1	-
Korez et al.[2][68]	93.1±2.0	-	-
Hammernik et al.[44]	93.0±4.0	-	-

Table 6.3: Quantitative results of CSI2014seg dataset.

evaluated their approach using a leave-one-out 3-fold cross validation.

6.2 LIMITATION

Besides the publicly accessible spine datasets, we also evaluate our method on two in-house spine CT datasets from the collaborated medical device company and hospital respectively. The manual annotation is provided/hold by the technicians and the doctors. The data is used only for evaluation and not part of the training set. As the datasets are internally accessible, we only compute the accuracy and inspect the results without comparing with other state of the art methods. A few limitations different from the ones listed in Sec. 5.5 are observed during the inspection.

EOS imaging test data From the collaborated medical device company, we obtain 3 test data. All of the three test data are scanning the full spine from C1 to L5 as shown in Figure 6.4. The first two subjects are elder patients, both with contrast agents injected in

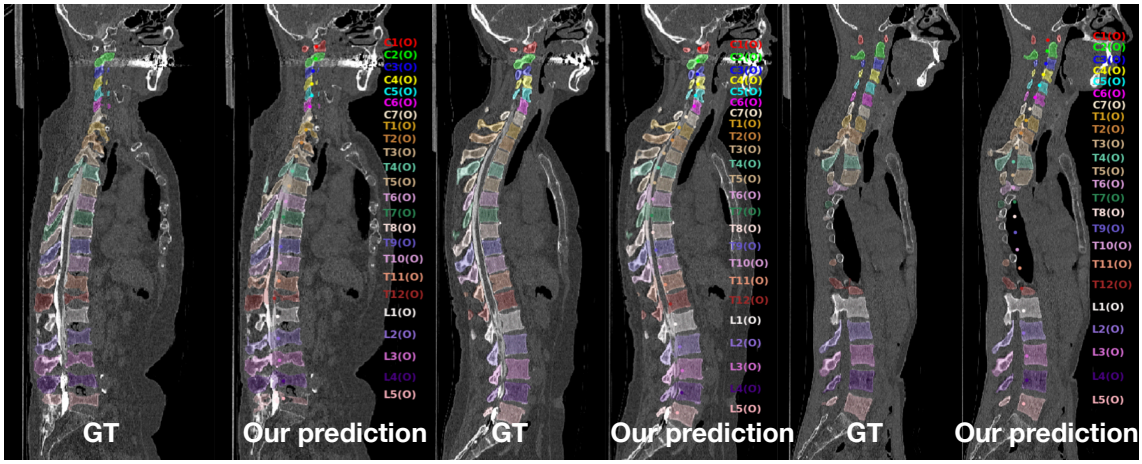


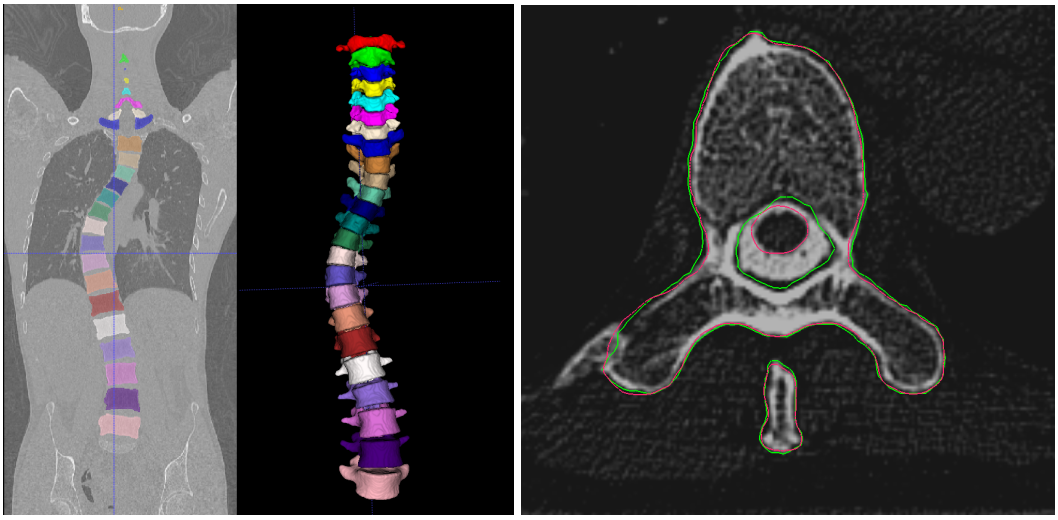
Figure 6.4: Results visualization of EOS imaging test data.

the spinal cords. For Dice scores (%) we obtain 85.88 ± 3.11 and 91.05 ± 1.68 respectively. For HD (mm) we obtain 7.33 ± 2.01 and 6.30 ± 2.69 respectively. The third subject is a young adult with scoliosis. We obtain Dice score (%) of 93.38 ± 2.00 and HD (mm) of 4.45 ± 4.86. The obtained accuracy indicates that our method generalizes well to unseen datasets and is able to provide adequate segmentation and identification.

By inspecting the results slice by slice, we found that the contrast agents that were injected in the patients' spinal cord are segmented as bone. As shown in Figure 6.5 (b), the green line represents the contour of GT segmentation and red line represents our prediction. The brightness caused by the contrast agent presents different and new features that the training set did not observe, so that our method is not robust to this contrast agent. Nevertheless, our method is well generalized to the scoliotic spine. Figure 6.5 (a) shows the segmented spine which presents a moderate curve ($25 - 40$ degrees). Thanks to the data augmentation strategy, our method is robust to the tilted vertebrae.

CHU test data We are blind to this test set in the way that the collaborated hospital took our packed implementation and tested our method on their own. We have no access to the test data and only received the statistics of the results from them. The test data consists of 675 CT scans associated with the manual annotations from the doctor. All the scans include the abdomino-pelvic cavity, where part of the spine is presented, but are not restricted to this field of view.

Among 675 CT scans, 20 scans did not provide any output because of the memory issue or non converged individual vertebra segmentor. The rest provided promising results where the accuracy can be found in Table 6.4. After obtaining the vertebrae segmentation and identification by running our code, the doctor further refined their initial manual annotation taking our prediction as reference. We see that with the refined ground truth, we obtain 93.6% Dice score, 3.7% higher than that with the initial annotation. If the 20 failures are excluded, the Dice score can reach 96.5%.



(a) Subject with scoliosis (b) Subject with contrast agent injected

Figure 6.5: Results visualization of special cases; (a) Vertebrae segmentation and identification of a full spine with scoliosis; (b) An axial slice of the vertebra. Green line is the GT segmentation, red line is our prediction. The contrast agent is segmented as bone by our method.

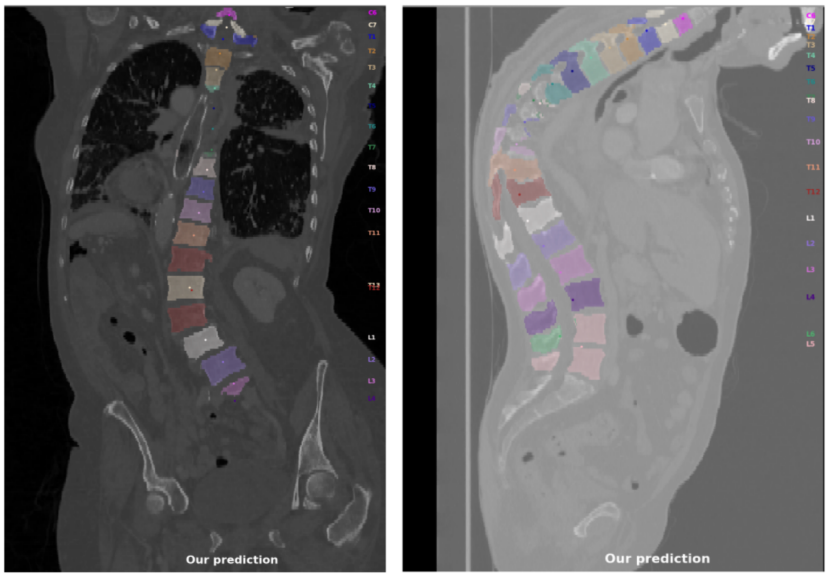
Dice score (%)	with the initial annotations	with the refined annotations
including the empty predictions	89.9	93.6
excluding the empty predictions	92.8	**96.5**

Table 6.4: Quantitative results of CHU test set.

A significant fact was reported that the doctor corrected their initial annotation by adding 25 L6. They are missed in the manual segmentation but detected by our method. It indicates the normality of sacralization/lumbarization in the population. In total around 10% L6 are found in the test set which is coherent with the statistics [21, 113, 65].

Apart from the 20 scans that provided empty output, among the 675 test scans, 23 scans obtained unqualified predictions. Most of them have errors in the L5/Sacrum area. Our method always tends to look for L6 in the field of view, it causes an issue that some S1 which are not separated from the sacrum are being well segmented and identified as L6. An improvement can be made towards to the graph that extra nodes can be added for modeling the sacrum.

A few failure cases are observed in abnormal subjects or the ones with severe deformities, such as severe scoliosis and kyphosis as shown in Fig. 6.6. The spine with severe kyphosis has up to 90 degrees curvature, which means the axial plane of the vertebrae is orthogonal to the axial plane of the CT volume. In our algorithm, we pre-process all the input scan into a same anatomical orientation and run the pipeline. The spine is allowed to have some rotations (50 degrees tolerance) as we augmented the training data. The trained network is not robust to the severe scoliosis where the vertebrae are largely



(a) Spine with scoliosis (b) Spine with kyphosis

Figure 6.6: Failures on severe scoliosis and kyphosis.

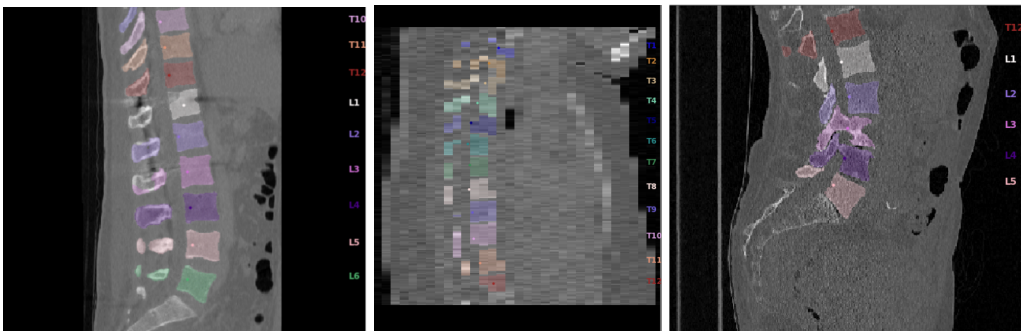
rotated. Towards this limitation, future work can improve the method by augmenting the training data with larger variance.

For the failure cases of the non converged individual vertebra segmentation, it either segments nothing at the detected location, or it segments something but lost it in the iterative location-segmentation refinement scheme. As we do not have access to the data, the exact reason of why the image distribution does not fit to the trained model is unknown. In principle, the CT scans even from different devices respect the hounsfield unit. The intensity range of the image should be covered by the training data. A hypothesis is that the contrast of the scan is affected so the extracted feature becomes different from the trained ones.

Overall, our method achieved considerable results on this large test set. For some extreme cases such as metal-inserted vertebrae, fractures vertebrae and very poor-quality scans, it also provided promising outputs as shown in Fig 6.7.

6.3 Conclusion

In this chapter, we thoroughly investigated the generalization ability and analyze the limitation of our method of segmenting and identifying vertebrae from CT images. We adopted both the publicly available datasets which we can compare our approach with other methods, and the in-house datasets that we do not have access to the data and their associated annotation. We observed that our method has a good generalization capability as it achieved superior or comparable results when comparing with the state of the art methods on the benchmark. However, the limitations are also remarked. They can



(a) Scan with metals (b) Scan with poor quality (c) Scan with fractured vertebrae

Figure 6.7: Qualitative results on extreme cases.

be summarized as: a) The identifications are with one label shift, mostly happened in the field of thoracolumbar or lumbar area; b) Our method is not robust to the contrast agent; c) Our method is not robust to the spine with severe scoliosis; d) Our method false identified S1 as L6; e) The individual vertebra segmentor can fail in some distorted CT scans. Considering these limitations, a list of improvements can be addressed in the future work. To solve a), the individual vertebra classification network can be task-oriented re-designed rather than using a simple backbone network. In this way, higher probability of each vertebra being classified into the correct class is expected, a more accurate global configuration is expected from the graph optimization. As for b) and c), more data can be augmented and included in the training set. For instance, the individual vertebra can be fully rotated during data augmentation. d) can be solved by including the sacrum in the graph. e) can be potentially improved by applying a more appropriate pre-processing than a standard normalization. Overall, our method is capable of providing promising results to the unseen CT scans while still needing some improvements regarding the corner cases.

CHAPTER 7

DISCUSSION AND CONCLUSION

7.1 SUMMARY

This thesis focus on the human spine and vertebrae in CT images. Firstly, we learned a statistical full spine surface model from partial observations using PPCA. In the training set, not a single full spine was observed. The model can faithfully capture the shape of the vertebrae and generalize to the left out data. Given an incomplete view of the spine, the model can predict the unseen vertebrae with a distance error under 3mm. Further, we investigated deep networks for spine segmentation. We explored both the spine binary segmentation and an end-to-end trained multi-vertebrae segmentation. We concluded that the spine binary segmentation is a better choice for the first step of a multi-stage vertebrae segmentation approach. In contrast, the straight forward multi-vertebrae segmentation lacks spatial consistency and a single network is less capable of performing the segmentation and identification tasks simultaneously. The end-to-end network is confused by the similar neighboring vertebrae. Nevertheless, the experiments show interesting results that the vertebra can be accurately categorized into anatomic groups rather than individual labels, which inspired us a hierarchical framework for individual vertebra classification. For the vertebrae identification, we use the vertebrae shape as input rather than the image. The quantitative results show that using the morphology of the vertebrae obtained comparable/higher performance than using the image itself. The benefit of adopting the vertebrae shape is modality invariant. The learned model can be easily transferred to other modalities such as MRI. We proposed an approach combing the local and global strategies for vertebrae identification. For each vertebra, we use a hierarchical framework. It first classifies the vertebra into an anatomic group, then predicts its individual label inside the group. The problem of classifying the vertebra into one of 24 classes is turned into classifying the vertebra into one of $7/12/5$ classes. The intra-class variance is largely reduced. We further proposed a graphical model to reason on the predicted individual labels in a global level. The graph enforces the anatomical consistency and it explicitly models the transitional vertebrae to well detect the presence of L6, T13 and the absence of T12. For the unified task of vertebrae segmentation, localization and identification, we

leveraged the shape statistics and the deep networks. The pathological cases such as fractured vertebrae and metal-inserted vertebrae, which are difficult to be learned from the limited data for their poor representation, can be well treated with the learned statistics. To be specific, the statistics of the vertebrae volume size and inter-vertebral distances are employed. The shape constraints are enforced in the iterative cycle of segmenting, localizing and identifying vertebrae. An anatomically consistent result is obtained. Otherwise the region of where an inconsistency raised is reported for diagnosis.

7.2 DISCUSSION

We presented an anatomic consistency cycle for vertebrae localization, segmentation and identification in arbitrary field of view CT images. The set of vertebrae locations, masks and labels is cycled through until an anatomical coherence is achieved. Compared with most existing methods, which use a chain of modules to perform the three tasks, the cycling process avoids the accumulated errors from the sequential blocks. Moreover, the three tasks cooperate and complement each other boosting the performance. In our method, the vertebra location and segmentation benefit and refine each other. The anchor location is used to segment the vertebra and it can in turn be computed from the segmented mask. A stable pair of vertebra location and segmentation is obtained through the iterative process. The iterative refinement scheme is inspired by Lessmann et al. [70], in their method a segmentation patch guides a sliding window until a complete vertebra is seen. The vertebra location and its corresponding mask are then used for identification. The identification in turn condition the vertebral statistics to locate individual vertebrae. The cycle unifies the three tasks and utilizes the anatomy, assuring the obtained set of locations, segmented masks and labels is anatomically consistent. We regard the proposed anatomic consistency cycle as a general framework for multi-label segmentation task in medical images, in the sense that the individual modules can be replaced. For instance, the segmentation or the identification network can be customized [94, 109]. We use the baseline networks such as U-net [98] and vgg [105], as designing a new architecture is not the interest of this work. The cycle with the defined consistency criteria can also be used as post processing to refine the result.

In the consistency cycle, we proposed to leverage the anatomical priors with deep neural networks. Generally a large amount of data benefit the supervised learning, low present samples are difficult to be learnt. In this work, we combine both strategies where the deep networks promise the learned outcome while the anatomical priors save the abnormal cases, e.g. pathological or transitional vertebrae. Fractured, compressed or other pathological vertebrae present non-standard shape, the hard constraints such as the appearance and image models do not apply. We chose to learn the statistics of the vertebra volume and the inter-vertebral distance. The volume quantifies a part of the vertebra shape, it is robust to morphological deviations. The inter-vertebral distance regularizes the vertebrae distributions. The two features are used as soft constraints to locate the individual vertebrae. The located vertebrae are then segmented by the individual segmentor. Through the iterative refinement process, either the location is stable with the segmentation mask

and added, or the mask is empty and the location is discarded. In the circumstance that the location added by the anatomic constraints is not consistent with the segmentor, we give the anatomy the priority and add the location. The hypothesis for this inconsistency is that the individual vertebra patch is highly distorted, e.g. there is a metal implant. The anatomical priors are additionally used as consistency criteria validating the final output is anatomically coherent. We report the inconsistent region in the result, available for the specialist to further diagnose.

We use a hierarchical approach for the individual vertebra classification task. A vertebra is classified through two stages, first the anatomic group is identified and then the individual label. Similar strategies are also found in [104, 87] where they first separate the spine into different anatomic regions, then segment and label the individual vertebrae. The two-stage method largely reduces the intra-class variances and improves the prediction accuracy. We employed the vertebrae morphology for the classification. Most existing works use the CT image as input to classify vertebrae with the intuition that the original scan provided more information. Our motivation was to investigate if the shape of the vertebrae itself is capable of being distinguished. We experimentally showed that the vertebrae shape is representable compared with the image patch, higher/comparable accuracy was achieved. The advantage of classifying vertebra from its shape is modality independent. The trained classifier is generalized and not restricted to one modality (e.g. CT or MRI) particularly. However, the prerequisite of using the shape is to segment the spine from the original scan. Vertebrae with incomplete or corrupted shape would thus affect the precision of the prediction. Even with the completely segmented vertebrae, the individual predictions may have errors. It motivated us to combine the local predictions with a global reasoning. The proposed graph enforces the consistency over individual predictions. It tolerates a certain number of inaccurate individual predictions and sorts out an optimal configuration. Another key feature of the proposed graph is that it explicitly models the transitional vertebrae, which affect $\sim 25\%$ of the general population but are limited present in the publicly available training set. With the graph, our method achieved the state-of-the-art performance in detecting the transitional vertebrae. It is worth noting that in the evaluation of the in-house dataset, 25 extra L6 were detected among 675 CT scans. Our method well handles the corner cases while maintaining a good performance on general data.

Overall, our method achieved the state-of-the-art performance on VerSe challenge benchmark and obtained comparable results on other publicly available datasets. On the unseen in-house datasets, our approach generalizes well and provided promising results. We experimentally showed that the method is robust to fractured, metal-inserted and compressed vertebrae, mild to moderate scoliosis, extreme poor-quality scans. Nevertheless, it still finds challenges in severe scoliotic spine and transitional vertebrae. A main failure of the method is the one-label-shift issue. We found that with the limited field of view, usually when only the thoracolumbar area is observed, the individual vertebra classification provides high confidence predicting one-offset labels. We use a baseline architecture vgg for the classification task. Further improvement can be made to explore on classification network optimization or data augmentation.

Considering the time and GPU cost of our method, inference times on VerSe20 public testset (103 CT scans) vary between 1.5 min to 248 min depending on the scan size, with a median runtime of 26 min. For the in-house dataset of 675 CT scans, the inference time varies between 7-90 min ,with a median runtime of 16.2 min. The time is heavily devoted to the spine binary segmentation, as the spine is segmented with intensely overlapped patches. It can be drastically improved by considering parallelization over patches. It also costs a lot when there are more than two loops in the anatomic consistency cycle. Nevertheless, the method is GPU efficient with no more than 3 GB occupation. The CPU memory is costly when the individual vertebrae volumes are processed and accumulated in the cycle.

7.3 FUTURE RESEARCH DIRECTIONS

From this thesis, there are several possible future research directions:

- In Chapter 4, we use a backbone architecture 3D *vgg16* for the individual vertebra classification. The network takes an individual vertebra as input and outputs the predicted class with associated probabilities. The backbone is not powerful enough for this specific task as seen in the evaluation of the failure cases in Chapter 5, the class is predicted with one offset with a high confidence. The network confuses the neighboring thoracolumbar vertebrae. It would be interesting to investigate if the backbone architecture can be optimized towards this task, such as multi-scale pyramid architecture. Moreover, contrast learning would be interesting to be explored to further distinguish between the similar identities for this circumstance.
- The graphical model we proposed for vertebrae identification considers 24 vertebrae with possible presence or absence of the transitional vertebrae. For future work, it can be extended to include the sacrum, femur, or other anatomy of the body. The same for the proposed anatomic consistency cycle that leverages the shape prior and deep networks, it can be applied to other body structures as the statistics of prior anatomic knowledge can be learned and enforced. To a local extent, each body structure can be segmented and identified using deep networks. For the global reasoning, the anatomic consistency can be enforced through the cycle.
- In this thesis, segmenting the vertebrae and identifying them from CT images is explored and achieves promising performance. The next step is to transfer the method to CBCT images where there is limited available annotation. Two main directions can be explored: one is to minimize the domain gap by mimicking the image appearance, which is to generate stylized CT image from CBCT image; The other one is to learn a transformation function in the feature space, the features that represent the spine in CBCT image can be mapped to the CT feature space regardless the image appearance. So that the segmentation map can be reconstructed from the feature space as well as the identity prediction.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

[1] Picuki: Vertebrae (2020). https://www.picuki.com/media/2227838139 5-84745065, 2020. [Online; accessed April-2020].

[2] Vertebral column. https://www.britannica.com/science/vertebral-column, 2022. [Online; accessed April-2022].

[3] Wikipedia: Vertebra. https://en.wikipedia.org/wiki/Vertebra, 2022. [Online; accessed April-2022].

[4] SM Al Arif, Michael Gundry, Karen Knapp, and Greg Slabaugh. Improving an active shape model with random classification forest for segmentation of cervical vertebrae. In *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, pages 3–15. Springer, 2016.

[5] SM Al Arif, Karen Knapp, and Greg Slabaugh. Shape-aware deep convolutional neural network for vertebrae segmentation. In *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pages 12–24. Springer, 2017.

[6] SM Masudur Rahman Al Arif, Karen Knapp, and Greg Slabaugh. Fully automatic cervical vertebrae segmentation framework for x-ray images. *Computer methods and programs in biomedicine*, 157:95–111, 2018.

[7] SM Masudur Rahman Al Arif, Karen Knapp, and Greg Slabaugh. Spnet: Shape prediction using a fully convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 430–439. Springer, 2018.

[8] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on pattern analysis and machine intelligence*, (5):698–700, 1987.

[9] APRN FNP-BC Aileen Lin Ashley Mauldin, MSN. Anatomical Position: What it is, significance, regions, planes and more. https://www.osmosis.org/answers/anatomical-position, 2022. [Online; accessed April-2022].

[10] Melih S Aslan, Asem Ali, Dongqing Chen, Ben Arnold, Aly A Farag, and Ping Xiang. 3d vertebrae segmentation using graph cuts with shape prior constraints. In *2010 IEEE International Conference on Image Processing*, pages 2193–2196. IEEE, 2010.

[11] Mohammed Benjelloun, Saïd Mahmoudi, and Fabian Lecron. A framework of vertebra segmentation using the active shape model-based approach. *International journal of biomedical imaging*, 2011, 2011.

[12] John H Bland and Dallas R Boushey. Anatomy and physiology of the cervical spine. In *Seminars in arthritis and rheumatism*, volume 20, pages 1–20. Elsevier, 1990.

[13] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[14] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 2300–2308, 2015.

[15] Matías N Bossa and Salvador Olmos. Multi-object statistical pose+ shape models. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1204–1207. IEEE, 2007.

[16] Paul A Bromiley, Eleni P Kariki, Judith E Adams, and Timothy F Cootes. Fully automatic localisation of vertebrae in ct images using random forest regression voting. In *International Workshop on Computational Methods and Clinical Applications for Spine Imaging*, pages 51–63. Springer, 2016.

[17] Philipp Bruners, Tobias Penzkofer, Markus Nagel, Robert Elfring, Nina Gronloh, Thomas Schmitz-Rode, Rolf W Günther, and Andreas H Mahnken. Electromagnetic tracking for ct-guided spine interventions: phantom, ex-vivo and in-vivo results. *European radiology*, 19(4):990–994, 2009.

[18] Yunliang Cai, Said Osman, Manas Sharma, Mark Landis, and Shuo Li. Multi-modality vertebra recognition in arbitrary views using 3d deformable hierarchical model. *IEEE transactions on medical imaging*, 34(8):1676–1693, 2015.

[19] JQ Campbell and AJ Petrella. Automated finite element modeling of the lumbar spine: using a statistical shape model to generate a virtual population of models. *Journal of biomechanics*, 49(13):2593–2599, 2016.

[20] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[21] John A Carrino, Paul D Campbell Jr, Dennis C Lin, William B Morrison, Mark E Schweitzer, Adam E Flanders, John Eng, and Alexander R Vaccaro. Effect of spinal

segment variants on numbering vertebral levels at lumbar mr imaging. *Radiology*, 259(1):196–202, 2011.

[22] Isaac Castro-Mateos, Jose M Pozo, Marco Pereañez, Karim Lekadir, Aron Lazary, and Alejandro F Frangi. Statistical interspace models (sims): application to robust 3d spine segmentation. *IEEE transactions on medical imaging*, 34(8):1663–1675, 2015.

[23] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 865–872, 2019.

[24] Di Chen, Yiwei Bai, Wenting Zhao, Sebastian Ament, John Gregoire, and Carla Gomes. Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In *International Conference on Machine Learning*, pages 1500–1509. PMLR, 2020.

[25] Hao Chen, Chiyao Shen, Jing Qin, Dong Ni, Lin Shi, Jack CY Cheng, and Pheng-Ann Heng. Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 515–522. Springer, 2015.

[26] Yizhi Chen, Yunhe Gao, Kang Li, Liang Zhao, and Jun Zhao. vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model. *IEEE transactions on medical imaging*, 39(2):387–399, 2019.

[27] Pengfei Cheng, Yusheng Yang, Huiqiang Yu, and Yongyi He. Automatic vertebrae localization and segmentation in ct with a two-stage dense-u-net. *Scientific Reports*, 11(1):1–13, 2021.

[28] Chengwen Chu, Daniel L Belavỳ, Gabriele Armbrecht, Martin Bansmann, Dieter Felsenberg, and Guoyan Zheng. Fully automatic localization and segmentation of 3d vertebral bodies from ct/mr images via a learning-based method. *PloS one*, 10(11):e0143327, 2015.

[29] Cheng-Hung Chuang, Chih-Yang Lin, Yuan-Yu Tsai, Zhen-You Lian, Hong-Xia Xie, Chih-Chao Hsu, and Chung-Lin Huang. Efficient triple output network for vertebral segmentation and identification. *IEEE Access*, 7:117978–117985, 2019.

[30] Patrick M Colletti, Herrick J Siegel, Mary Y Woo, Howard Y Young, and Michael R Terk. The impact on treatment planning of mri of the spine in patients suspected of vertebral metastasis: an efficacy study. *Computerized medical imaging and graphics*, 20(3):159–162, 1996.

[31] Timothy F Cootes and Christopher J Taylor. Combining point distribution models with shape models based on finite element analysis. *Image and Vision Computing*, 13(5):403–409, 1995.

[32] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[33] Timothy F Cootes, Christopher J Taylor, and Andreas Lanitis. Active shape models: Evaluation of a multi-resolution method for improving image search. In *BMVC*, volume 1, pages 327–336. Citeseer, 1994.

[34] Uwe Dick, Peter Haider, and Tobias Scheffer. Learning from incomplete data with infinite imputations. In *Proceedings of the 25th international conference on Machine learning*, pages 232–239, 2008.

[35] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[36] J Dvorak, D Froehlich, L Penning, H Baumgartner, and MM Panjabi. Functional radiographic diagnosis of the cervical spine: flexion/extension. *Spine*, 13(7):748–755, 1988.

[37] Todd M Emch and Michael T Modic. Imaging of lumbar degenerative disk disease: history and current state. *Skeletal radiology*, 40(9):1175–1189, 2011.

[38] Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational statistics & data analysis*, 52(3):1694–1711, 2008.

[39] Daniel Forsberg, Claes Lundström, Mats Andersson, Ludvig Vavruch, Hans Tropp, and Hans Knutsson. Fully automatic measurements of axial vertebral rotation for assessment of spinal deformity in idiopathic scoliosis. *Physics in Medicine & Biology*, 58(6):1775, 2013.

[40] Subarna Ghosh, Alomari Raja'S, Vipin Chaudhary, and Gurmeet Dhillon. Automatic lumbar vertebra segmentation from clinical ct for wedge compression fracture diagnosis. In *Medical Imaging 2011: Computer-Aided Diagnosis*, volume 7963, pages 21–29. SPIE, 2011.

[41] Ben Glocker, Johannes Feulner, Antonio Criminisi, David R Haynor, and Ender Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 590–598. Springer, 2012.

[42] Ben Glocker, Darko Zikic, Ender Konukoglu, David R Haynor, and Antonio Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In *International conference on medical image computing and computer-assisted intervention*, pages 262–270. Springer, 2013.

[43] Benjamin Gutierrez, Diana Mateus, Ehab Shiban, Bernhard Meyer, Jens Lehmberg, and Nassir Navab. A sparse approach to build shape models with routine clinical data. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 258–261. IEEE, 2014.

[44] Kerstin Hammernik, Thomas Ebner, Darko Stern, Martin Urschler, and Thomas Pock. Vertebrae segmentation in 3d ct images based on a variational framework. In *Recent advances in computational methods and clinical applications for spine imaging*, pages 227–233. Springer, 2015.

[45] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009.

[46] Justin FM Hollenbeck, Christopher M Cain, Jill A Fattor, Paul J Rullkoetter, and Peter J Laz. Statistical shape modeling characterizes three-dimensional shape and alignment variability in the lumbar spine. *Journal of biomechanics*, 69:146–155, 2018.

[47] Benjamin Howe, Arunkumar Gururajan, Hamed Sari-Sarraf, and L Rodney Long. Hierarchical segmentation of cervical and lumbar vertebrae using a customized generalized hough transform and extensions to active appearance models. In *6th IEEE Southwest Symposium on Image Analysis and Interpretation, 2004.*, pages 182–186. IEEE, 2004.

[48] Szu-Hao Huang, Yi-Hong Chu, Shang-Hong Lai, and Carol L Novak. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal mri. *IEEE transactions on medical imaging*, 28(10):1595–1605, 2009.

[49] Bulat Ibragimov, Robert Korez, Boštjan Likar, Franjo Pernuš, Lei Xing, and Tomaž Vrtovec. Segmentation of pathological structures by landmark-assisted deformable models. *IEEE transactions on medical imaging*, 36(7):1457–1469, 2017.

[50] Bulat Ibragimov, Boštjan Likar, Franjo Pernuš, and Tomaž Vrtovec. Shape representation for efficient landmark-based segmentation in 3-d. *IEEE transactions on medical imaging*, 33(4):861–874, 2014.

[51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[52] Roman Jakubicek, Jiri Chmelik, Jiri Jan, Petr Ourednicek, Lukas Lambert, and Giampaolo Gavelli. Learning–based vertebra localization and labeling in 3d ct data of possibly incomplete and pathological spines. *Computer methods and programs in biomedicine*, 183:105081, 2020.

[53] Rens Janssens, Guodong Zeng, and Guoyan Zheng. Fully automatic segmentation of lumbar vertebrae from ct images using cascaded 3d fully convolutional networks. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 893–897. IEEE, 2018.

[54] Samuel Kadoury, Hubert Labelle, and Nikos Paragios. Automatic inference of articulated spine models in ct images using high-order markov random fields. *Medical image analysis*, 15(4):426–437, 2011.

[55] Samuel Kadoury, Hubert Labelle, and Nikos Paragios. Spine segmentation in medical images using manifold embeddings and higher-order mrfs. *IEEE transactions on medical imaging*, 32(7):1227–1238, 2013.

[56] B Michael Kelm, Michael Wels, S Kevin Zhou, Sascha Seifert, Michael Suehling, Yefeng Zheng, and Dorin Comaniciu. Spine detection in ct and mr using iterated marginal space learning. *Medical image analysis*, 17(8):1283–1292, 2013.

[57] Pulkit Khandelwal, D Louis Collins, and Kaleem Siddiqi. Spine and individual vertebrae segmentation in computed tomography images using geometric flows and shape priors. *Frontiers in Computer Science*, page 66, 2021.

[58] Kang Cheol Kim, Hyun Cheol Cho, Tae Jun Jang, Jong Mun Choi, and Jin Keun Seo. Automatic detection and segmentation of lumbar vertebrae from x-ray images for compression fracture evaluation. *Computer Methods and Programs in Biomedicine*, 200:105833, 2021.

[59] Yiebin Kim and Dongsung Kim. A fully automatic vertebra segmentation method using 3d deformable fences. *Computerized Medical Imaging and Graphics*, 33(5):343–352, 2009.

[60] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[61] Matthias Kirschner, Meike Becker, and Stefan Wesarg. 3d active shape model segmentation with nonlinear shape priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 492–499. Springer, 2011.

[62] Tobias Klinder, Jörn Ostermann, Matthias Ehm, Astrid Franz, Reinhard Kneser, and Cristian Lorenz. Automated model-based vertebra detection, identification, and segmentation in ct images. *Medical image analysis*, 13(3):471–482, 2009.

[63] Dejan Knez, Boštjan Likar, Franjo Pernuš, and Tomaž Vrtovec. Computer-assisted screw size and insertion trajectory planning for pedicle screw placement surgery. *IEEE transactions on medical imaging*, 35(6):1420–1430, 2016.

[64] Martin Kolařík, Radim Burget, Václav Uher, Kamil Říha, and Malay Kishore Dutta. Optimized high resolution 3d dense-u-net network for brain and spine segmentation. *Applied Sciences*, 9(3):404, 2019.

[65] GP Konin and DM Walz. Lumbosacral transitional vertebrae: classification, imaging findings, and clinical relevance. *American Journal of Neuroradiology*, 31(10):1778–1786, 2010.

[66] Soontharee Koompairojn, Kien A Hua, and Chutima Bhadrakom. Automatic classification system for lumbar spine x-ray images. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 213–218. IEEE, 2006.

[67] Robert Korez, Bulat Ibragimov, Boštjan Likar, Franjo Pernuš, and Tomaž Vrtovec. A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. *IEEE transactions on medical imaging*, 34(8):1649–1662, 2015.

[68] Robert Korez, Bulat Ibragimov, Boštjan Likar, Franjo Pernuš, and Tomaž Vrtovec. Interpolation-based shape-constrained deformable model approach for segmentation of vertebrae from ct spine images. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pages 235–240. Springer, 2015.

[69] Mohamed Amine Larhmam, Mohammed Benjelloun, and Saïd Mahmoudi. Vertebra identification using template matching modelmp and k-means clustering. *International journal of computer assisted radiology and surgery*, 9(2):177–187, 2014.

[70] Nikolas Lessmann, Bram Van Ginneken, Pim A De Jong, and Ivana Išgum. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical image analysis*, 53:142–155, 2019.

[71] Nikolas Lessmann, Bram van Ginneken, and Ivana Išgum. Iterative convolutional neural networks for automatic vertebra identification and segmentation in ct images. In *Medical Imaging 2018: Image Processing*, volume 10574, page 1057408. International Society for Optics and Photonics, 2018.

[72] 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, 2018.

[73] Poay Hoon Lim, Ulas Bagci, and Li Bai. Introducing willmore flow into level set segmentation of spinal vertebrae. *IEEE Transactions on Biomedical Engineering*, 60(1):115–122, 2012.

[74] Claudia Lindner, Paul A Bromiley, Mircea C Ionita, and Tim F Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1862–1874, 2014.

[75] Yingzi Liu, Yang Lei, Tonghe Wang, Yabo Fu, Xiangyang Tang, Walter J Curran, Tian Liu, Pretesh Patel, and Xiaofeng Yang. Cbct-based synthetic ct generation using deep-attention cyclegan for pancreatic adaptive radiotherapy. *Medical physics*, 47(6):2472–2483, 2020.

[76] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020.

[77] L Rodney Long and George R Thoma. Use of shape models to search digitized spine x-rays. In *Proceedings 13th IEEE Symposium on Computer-Based Medical Systems. CBMS 2000*, pages 255–260. IEEE, 2000.

[78] M.J. Loper. Chumpy is a python-based framework designed to handle the auto- differentiation problem. `htttps://pypi.python.org/pypi/chumpy`, 2015. [Online; accessed Feb-2020].

[79] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[80] Cristian Lorenz and Nils Krahnstöver. Generation of point-based 3d statistical shape models for anatomical objects. *Computer vision and image understanding*, 77(2):175–191, 2000.

[81] Marcel Lüthi, Thomas Albrecht, and Thomas Vetter. Building shape models from lousy data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1–8. Springer, 2009.

[82] Jingting Ma, Anqi Wang, Feng Lin, Stefan Wesarg, and Marius Erdt. A novel robust kernel principal component analysis for nonlinear statistical shape modeling from erroneous data. *Computerized Medical Imaging and Graphics*, 77:101638, 2019.

[83] Jun Ma and Le Lu. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. *Computer Vision and Image Understanding*, 117(9):1072–1083, 2013.

[84] Alexander Oliver Mader, Cristian Lorenz, Martin Bergtholdt, Jens von Berg, Hauke Schramm, Jan Modersitzki, and Carsten Meyer. Detection and localization of spatially correlated point landmarks in medical images using an automatically learned conditional random field. *Computer Vision and Image Understanding*, 176:45–53, 2018.

[85] Alexander Oliver Mader, Cristian Lorenz, Jens von Berg, and Carsten Meyer. Automatically localizing a large set of spatially correlated key points: a case study in spine imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–392. Springer, 2019.

[86] Dennis Madsen, Jonathan Aellen, Andreas Morel-Forster, Thomas Vetter, and Marcel Lüthi. Learning shape priors from pieces. In *International Workshop on Shape in Medical Imaging*, pages 30–43. Springer, 2020.

[87] Naoto Masuzawa, Yoshiro Kitamura, Keigo Nakamura, Satoshi Iizuka, and Edgar Simo-Serra. Automatic segmentation, localization, and identification of vertebrae in 3d ct images using cascaded convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 681–690. Springer, 2020.

[88] James McCouat and Ben Glocker. Vertebrae detection and localization in ct with two-stage cnns and dense annotations. *arXiv preprint arXiv:1910.05911*, 2019.

[89] Hengameh Mirzaalian, Michael Wels, Tobias Heimann, B Michael Kelm, and Michael Suehling. Fast and robust 3d vertebra segmentation using statistical shape models. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3379–3382. IEEE, 2013.

[90] Lorenzo Nardo, Hamza Alizai, Warapat Virayavanich, Felix Liu, Alexandra Hernandez, John A Lynch, Michael C Nevitt, Charles E McCulloch, Nancy E Lane, and Thomas M Link. Lumbosacral transitional vertebrae: association with low back pain. *Radiology*, 265(2):497–503, 2012.

[91] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[92] Shumao Pang, Chunlan Pang, Zhihai Su, Liyan Lin, Lei Zhao, Yangfan Chen, Yujia Zhou, Hai Lu, and Qianjin Feng. Dgmsnet: Spine segmentation for mr image by a detection-guided mixed-supervised segmentation network. *Medical image analysis*, 75:102261, 2022.

[93] P Papin, Hubert Labelle, Sébastien Delorme, C-E Aubin, Jacques A de Guise, and Jean Dansereau. Long-term three-dimensional changes of the spine after posterior spinal instrumentation and fusion in adolescent idiopathic scoliosis. *European Spine Journal*, 8(1):16–21, 1999.

[94] Christian Payer, Darko Stern, Horst Bischof, and Martin Urschler. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. In *VISIGRAPP (5: VISAPP)*, pages 124–133, 2020.

[95] Chunli Qin, Demin Yao, Han Zhuang, Hui Wang, Yonghong Shi, and Zhijian Song. Residual block-based multi-label classification and localization network with integral regression for vertebrae labeling. *arXiv preprint arXiv:2001.00170*, 2020.

[96] Kumar T Rajamani, Martin A Styner, Haydar Talib, Guoyan Zheng, Lutz P Nolte, and Miguel A González Ballester. Statistical deformable bone models for robust 3d surface extrapolation from sparse data. *Medical Image Analysis*, 11(2):99–109, 2007.

[97] Abtin Rasoulian, Robert Rohling, and Purang Abolmaesumi. Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model. *IEEE transactions on medical imaging*, 32(10):1890–1900, 2013.

[98] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[99] Silvia Ruiz-España, Juan Domingo, Antonio Díaz-Parra, Esther Dura, Víctor D'Ocón-Alcañiz, Estanislao Arana, and David Moratal. Automatic segmentation

of the spine by means of a probabilistic atlas with a special focus on ribs suppression. *Medical physics*, 44(9):4695–4707, 2017.

[100] 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. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 122–133. Springer, 2007.

[101] Sascha Seifert, Adrian Barbu, S Kevin Zhou, David Liu, Johannes Feulner, Martin Huber, Michael Suehling, Alexander Cavallaro, and Dorin Comaniciu. Hierarchical parsing and semantic navigation of full body ct data. In *Medical Imaging 2009: Image Processing*, volume 7259, page 725902. International Society for Optics and Photonics, 2009.

[102] Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021.

[103] Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentinitsch, Jan S Kirschke, and Bjoern H Menze. Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 649–657. Springer, 2018.

[104] Anjany Sekuboyina, Alexander Valentinitsch, Jan S Kirschke, and Bjoern H Menze. A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets. *arXiv preprint arXiv:1703.04347*, 2017.

[105] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[106] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[107] Roger Stevens. Gray's anatomy for students, 2006.

[108] IAF Stokes. Three-dimensional terminology of spinal deformity. a report presented to the research society working group on 3-d terminology of spinal deformity. *Spine*, 19(2):236–248, 1994.

[109] Rong Tao, Wenyong Liu, and Guoyan Zheng. Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine cts via 3d transformers. *Medical Image Analysis*, 75:102258, 2022.

[110] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[111] Thorsten Tjardes, Sven Shafizadeh, Dieter Rixen, Thomas Paffrath, Bertil Bouillon, Eva S Steinhausen, and Holger Baethis. Image-guided spine surgery: state of the art and future directions. *European spine journal*, 19(1):25–45, 2010.

[112] Andy Tsai, Anthony Yezzi, William Wells, Clare Tempany, Dewey Tucker, Ayres Fan, W Eric Grimson, and Alan Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *IEEE transactions on medical imaging*, 22(2):137–154, 2003.

[113] Demet Uçar, Bekir Yavuz Uçar, Yahya Coşar, Kurtuluş Emrem, Gürkan Gümüşsuyu, Serhat Mutlu, Burcu Mutlu, Mehmet Akif Çaçan, Yılmaz Mertsoy, and Hatice Gümüş. Retrospective cohort study of the prevalence of lumbosacral transitional vertebra in a wide and well-represented population. *Arthritis*, 2013, 2013.

[114] Malinda Vania, Dawit Mureja, and Deukhee Lee. Automatic spine segmentation from ct images using convolutional neural network via redundant generation of class labels. *Journal of Computational Design and Engineering*, 6(2):224–232, 2019.

[115] Jakob Verbeek. Probabilistic PCA python implementation. `https://github.com/shergreen/pyppca`, 2019. [Online; accessed Feb-2020].

[116] Veerle Vijvermans, Guy Fabry, and Jos Nijs. Factors determining the final outcome of treatment of idiopathic scoliosis with the boston brace: a longitudinal study. *Journal of Pediatric Orthopaedics B*, 13(3):143–149, 2004.

[117] Fakai Wang, Kang Zheng, Le Lu, Jing Xiao, Min Wu, and Shun Miao. Automatic vertebra localization and identification in ct by spine rectification and anatomically-constrained optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5280–5288, 2021.

[118] William Whitehead, Steven Moran, Bilwaj Gaonkar, Luke Macyszyn, and Subramanian Iyer. A deep learning approach to spine segmentation using a feed-forward chain of pixel-wise convolutional networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 868–871. IEEE, 2018.

[119] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. A convolutional approach to vertebrae detection and labelling in whole spine mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 712–722. Springer, 2020.

[120] Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

[121] Dong Yang, Tao Xiong, Daguang Xu, Qiangui Huang, David Liu, S Kevin Zhou, Zhoubing Xu, JinHyeong Park, Mingqing Chen, Trac D Tran, et al. Automatic vertebra labeling in large-scale 3d ct using deep image-to-image network with message

passing and sparsity regularization. In *International conference on information processing in medical imaging*, pages 633–644. Springer, 2017.

[122] Jianhua Yao, Joseph E Burns, Hector Munoz, and Ronald M Summers. Detection of vertebral body fractures based on cortical shell unwrapping. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 509–516. Springer, 2012.

[123] Yang C Yuan. Multiple imputation for missing data: Concepts and new development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, volume 267, 2000.

[124] Yiqiang Zhan, Dewan Maneesh, Martin Harder, and Xiang Sean Zhou. Robust mr spine detection using hierarchical learning and local articulated model. In *International conference on medical image computing and computer-assisted intervention*, pages 141–148. Springer, 2012.

[125] Shaoting Zhang, Yiqiang Zhan, Maneesh Dewan, Junzhou Huang, Dimitris N Metaxas, and Xiang Sean Zhou. Towards robust and effective shape modeling: Sparse shape composition. *Medical image analysis*, 16(1):265–277, 2012.

[126] Rongjian Zhao, Buyue Qian, Xianli Zhang, Yang Li, Rong Wei, Yang Liu, and Yinggang Pan. Rethinking dice loss for medical image segmentation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 851–860. IEEE, 2020.