TUM School of Computation, Information and Technology
Technische Universität München

Image Based Personalized Bone Therapy Planning using Deep Learning

Amirhossein Bayat

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Cristina Piazza

Prüfende der Dissertation:
1. Prof. Dr. Björn Menze
2. Prof. Dr. Marc-André Weber

Die Dissertation wurde am 29.11.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 08.11.2022 angenommen.

Abstract

Devising data-driven solutions to patient-specific biomechanical modeling or implant design facilitates the personalization of therapy planning and the prediction of surgery outcomes. Leveraging the recent advancements in deep learning and ever-increasing volumes of medical images being generated every day will improve the data-driven approaches in this regard. Individualized therapy planning embraces a broad range of applications and topics. In this publication-based dissertation, we aim for personalized anatomy modeling using deep learning and medical imaging. We explore the ideas of generating 3D biomechanical models for the spine and cranial implant design.

The balance of the spine is an essential factor in the development of degeneration, pain, and the outcome of spinal surgery. It should be analyzed in an upright, standing position to ensure physiological loading conditions and visualize load-dependent deformations. Despite the complex 3D shape of the spine, this analysis is currently conducted using 2D sagittal balance radiographs. All frequently used 3D imaging techniques require the patient to be scanned in a prone position.

We present a fully automatic method based on a deep neural network (DNN) to reconstruct the 3D shape of the spine from orthogonal 2D radiographs and vertebral centroid annotations. We train the DNN model on synthetic data, successfully deploy it on real-world clinical data, and reconstruct the 3D spinal pose in an upright standing position.

The proposed model for 3D reconstruction of the spine requires vertebral centroid annotation as one of the inputs. To automate the entire process and avoid manual work, we introduce another DNN model for accurate vertebrae localization and labeling in sagittal and coronal radiographs and to obtain the centroid annotations for the former model.

Automatic patient-specific implant designing is an understudied research area. The possibility of automatically designing patient-specific implants could reduce the surgery time and facilitates surgery planning. We explore the idea of designing the patient-specific implants from CT scans of defective skulls. We introduce a DNN model to design the 3D cranial implants. In the proposed method, the shape reconstruction is performed in two steps due to memory constraints. The 3D shape is reconstructed in low resolution first, and then another DNN model refines the results from the first step in 2D.

Zusammenfassung

Die Entwicklung datengesteuerter Lösungen für die patientenspezifische biomechanische Modellierung oder das Implantatdesign erleichtert die Personalisierung der Therapieplanung und die Vorhersage von Operationsergebnissen. Die Nutzung der jüngsten Fortschritte im Bereich Deep Learning und der ständig wachsenden Menge an medizinischen Bildern, die täglich generiert werden, wird die datengesteuerten Ansätze in dieser Hinsicht verbessern. Die individualisierte Therapieplanung umfasst ein breites Spektrum an Anwendungen und Themen. In dieser publikationsbasierten Dissertation streben wir eine personalisierte Anatomiemodellierung mithilfe von Deep Learning und medizinischer Bildgebung an. Wir erforschen die Ideen zur Generierung von biomechanischen 3D-Modellen für die Wirbelsäule und das Design von Schädelimplantaten.

Das Gleichgewicht der Wirbelsäule ist ein wesentlicher Faktor für die Entwicklung von Degeneration, Schmerzen und das Ergebnis von Wirbelsäulenoperationen. Sie sollte in einer aufrechten, stehenden Position analysiert werden, um physiologische Belastungsbedingungen zu gewährleisten und belastungsabhängige Verformungen zu visualisieren. Trotz der komplexen 3D-Form der Wirbelsäule wird diese Analyse derzeit mit 2D-Sagittal-Balance-Röntgenaufnahmen durchgeführt. Alle häufig verwendeten 3D-Aufnahmeverfahren erfordern, dass der Patient in liegender Position aufgenommen wird.

Wir stellen eine vollautomatische Methode vor, die auf einem tiefen neuronalen Netzwerk (DNN) basiert, um die 3D-Form der Wirbelsäule aus orthogonalen 2D-Röntgenbildern und Wirbelschwerpunktbeschriftungen zu rekonstruieren. Wir trainieren das DNN-Modell auf synthetischen Daten, setzen es erfolgreich auf realen klinischen Daten ein und rekonstruieren die 3D-Wirbelsäulenhaltung in einer aufrechten, stehenden Position.

Das vorgeschlagene Modell für die 3D-Rekonstruktion der Wirbelsäule benötigt die Annotation der Wirbelschwerpunkte als eine der Eingaben. Um den gesamten Prozess zu automatisieren und manuelle Arbeit zu vermeiden, führen wir ein weiteres DNN-Modell zur genauen Lokalisierung und Beschriftung von Wirbeln in sagittalen und koronalen Röntgenbildern ein, um die Schwerpunktbeschriftungen für das erste Modell zu erhalten.

Automatisches patientenspezifisches Implantatdesign ist ein wenig untersuchter Forschungsbereich. Die Möglichkeit, patientenspezifische Implantate automatisch zu entwerfen, könnte die Operationszeit reduzieren und die Operationsplanung erleichtern. Wir untersuchen die Idee, patientenspezifische Implantate aus CT-Scans von defekten Schädeln zu entwerfen. Wir führen ein DNN-Modell zum Design der 3D-Schädelimplantate ein. In der vorgeschlagenen Methode wird die Formrekonstruktion aufgrund von Speicherbeschränkungen in zwei Schritten durchgeführt. Die 3D-Form wird zunächst in niedriger Auflösung rekonstruiert, und dann verfeinert ein weiteres DNN-Modell die Ergebnisse aus dem ersten Schritt in 2D.

Acknowledgements

I cannot thank enough my supervisors, professor Menze from the Computer Science department and professor Kirschke from the Neuroradiology department of Klinikum recht der Isar, for all the support and for giving me the opportunity to research very exciting and challenging topics. I had the chance to explore different research areas and enjoy learning something new every day. Working in close collaboration with brilliant and hard-working people from all over the world taught me teamwork and gave me new ideas. I had a unique experience working with the university hospital. I could conduct research on real-world problems, be introduced to neuroradiologists' perspectives on research, and be in the clinical workspace. I want to thank all my colleagues and friends for inspiring and supporting me. Special thanks to (alphabetical order) Adrian Kilian, Akram Bayat, Alexader Valentinisch, Anjany Sekuboyina, Carolin Pirkl, Danielle Pace, Dhritiman Das, Diana Waldmannstetter, Esther Alberts, Felix Hofmann, Fernando Navarro, Florian Kofler, Giles Tetteh, Hanwool Park, Hongwei Li, Ivan Ezhov, Jana Lipkova, Johannes Paetzhold, John LaMaster, Judith Zimmermann, Jürgen Lichtenstein, Lina Xu, Malek El Husseini, Oguz Oztoprak, Oliver Schoppe, Rami Al-Maskari, Reihaneh Torkzadehmahani, Reza Nasirigerdeh, Sebiastian Endt, Suprosanna Shit, Timo Löhr, Tanja Lerchl, Xiaobin Hu, Yu Zhao, Yusuf Yilmaz.

Last, but not least, I would like to express my sincere gratitude to my family, whose love and support has been an incredible source of strength throughout my PhD journey. Thank you for your unwavering belief in me, for always being there for me, and for encouraging me to stay focused and determined. Your unconditional love and encouragement have been invaluable, and I could not have achieved this milestone without you.

Contents

List of Figures

List of Tables

Acronyms

AP	Anteroposterior.
B2V	Body to Vertebra.
C2V	Centroid to Vertebra.
CAD	Computer-aided Design.
CBCT	Cone Beam Computed Tomography.
CNN	Convolutional Neural Network.
CT	Computed Tomography.
DL	Deep Learning.
DNN	Deep Neural Network.
DOF	Degrees of Freedom.
DRR	Digitally Reconstructed Radiograph.
FCN	Fully Convolutional Neural Network.
FEM	Finite Element Modeling.
GAN	Generative Adversarial Network.
GPU	Graphics Processing Unit.
HU	Hounsfield units.
MBS	Multibody System Simulation.
MIP	Maximum Intensity Projection.
MLP	Multilayer Perceptron.
MRI	Magnetic Resonance Imaging.
NN	Nearest Neighborhood.
PCA	Principle Component Analysis.
PET	Positron Emission Tomography.
PSI	Patient Specific Implant.
RC	Residual Corrector.
ROI	Region of Interest.

SSM	Statistical Shape Model.
STN	Spatial Transformer Networks.
TPS	Thin-plate splines.
V2V	Vertebra to Vertebra.
VAE	Variational Autoencoder.
VOI	Vertebra of Interest.

1 Introduction

The continuous improvement in medical imaging made it essential for the diagnosis and treatment of patients. Due to ever-increasing volumes of complex data generated by imaging techniques, introducing new image analysis methods is needed to assist medical doctors. Nowadays, machine learning provides such methods for helping doctors in diagnosis, prognosis, and therapy planning.

Patient-specific biomechanical models enable us to personalize the therapy planning and surgery outcome prediction. In this publication-based dissertation, we aim at designing personalized models of spine and also cranial implant using deep learning and medical imaging. The primary objective of spine-related sections is to devise data-driven solutions for individualized therapy planning in back pain patients. After that, we study patient-specific implant design based on image data. We will leverage deep learning and image processing methodologies to propose new solutions for 3D spine modeling applicable in biomechanical analysis and 3D cranial implant design.

1.1 Contributions

This dissertation introduces different methods for processing 2D/3D images, shape reconstruction, and completion using deep learning. We applied the proposed methods in different applications, including spinal images and skull scans. More specifically, our contributions belong to the following fields:

- Developing a data processing pipeline, facilitating the medical image annotation, segmentation. This set of tools include DRR generation from CT scan, vertebral subregion segmentation, vertebral landmark detection. Generating the input to biomechanical analysis software like Simpack automatically.

- A deep neural network for localization and labeling vertebrae on radiographs. We introduce a residual coordinated corrector module on top of a Fully Convolutional Neural Network (FCN) to enforce the spinal shape and regress the vertebral centroid coordinates and labels.

- A deep neural network for inferring the 3D standing spinal posture from 2D radiographs. This network fuses the shape information from orthogonal 2D radiographs and reconstructs the 3D shape. We proposed a training scheme on synthetic data and successfully deployed the trained model on clinical radiographs due to lack of training data.

- A deep neural network for predicting the cranial implants given the defective skull. This network predicts the 3D shapes in low resolution and refines the prediction in 2D.

1.2 Outline of thesis

This thesis comprises three parts. In the beginning, we provide background information to the reader about concepts essential to this thesis's comprehension. Next, we present our works on spine. Finally, we introduce our approach for patient specific cranial implant design.

- Chapter 2 reviews the medical and technical background related to patient-specific bone modeling, including spine anatomy and spinal disorders in general, describes different medical imaging techniques applied for studying the spine and related analysis methods. This chapter also introduces the data processing pipelines developed for the works introduced in subsequent chapters.

- Chapter 3 introduces a deep learning method for vertebrae localization and labeling in spinal radiographs. Detecting and labeling vertebra in the radiograph is challenging due to the large field of view, heavy tissue overlay, and noise. Most of the works for vertebra labeling were on CT images; our work was one the first methods proposed for radiographs. We proposed a new architecture, robust to occlusion by enforcing the spinal shape.

- Chapter 4 presents a new deep learning based approach for synthesizing 3D shapes from 2D orthogonal images. An upright spinal pose (i.e., standing) under natural weight-bearing is crucial for biomechanical analysis. 3D volumetric imaging modalities (e.g. CT and Magnetic Resonance Imaging (MRI)) are performed in patients lying down. On the other hand, radiographs are captured in an upright pose but result in 2D projections. This work aims to integrate the two realms, i.e., it combines the upright spinal curvature from radiographs with the 3D vertebral shape from CT imaging for synthesizing an upright 3D model of the spine, loaded naturally. Specifically, we propose a novel neural network architecture working vertebra-wise, termed *TransVert*, which takes orthogonal 2D radiographs and infers the spine's 3D posture. We apply this model specifically to spinal radiographs. Applying this model to 2D radiographs enables us to generate 3D standing spinal postures applicable in biomechanical analysis of the spine. Training this model was a challenging task since the 3D ground-truths for radiographs are not available. Thus, we trained the model on synthetic data by simulating radiographs from CT images. However, the spinal CT scans available in the hospital are usually not the full field of view to decrease the scan size and save more disk space. Thus, the resulting 2D digitally reconstructed radiographs from CT images were not similar enough to real radiographs. We obtained large-scale datasets from lung scans where the lungs and ribs were available in the scan and created our synthetic dataset.

- Chapter 5 introduces a novel approach for designing patient-specific cranial implants using deep neural networks. Due to Graphics Processing Unit (GPU) memory limitations, we devised a solution composed of two sub-networks functioning on 3D and 2D images with low and high resolutions. Our model performs on 3D data in low resolution since a 2D model lacks a holistic 3D view of both the defective and healthy skulls. The first sub-network is designed to complete the shape of the downsampled defective skull. The second sub-network upsamples the reconstructed shape slice-wise. We train both the 3D and 2D networks in tandem in an end-to-end fashion, with a custom hierarchical loss function. Our proposed solution accurately predicts a high-resolution 3D implant in the challenge test case in terms of dice-score and the Hausdorff distance.
- Chapter 6 concludes the thesis by discussing the work presented in the following and suggesting directions for future work.

2 Background

In this chapter, we present the technical and medical background related to this thesis. We talk about the imaging modalities we used in this thesis, primarily radiographs and CT scans. We review the prior works on patient-specific therapy planning, mainly focused on the spine. We also review the related works on patient-specific implant design. Finally, we introduce the tools we developed to process the spinal scans and use them in our upcoming chapters.

2.1 Medical imaging modalities

Radiography and CT imaging demonstrate the x-ray absorption in human tissue, depending on the atoms or molecules' linear attenuation coefficient in the body tissue.

2.1.1 Radiograph

In spinal radiographs, the patient is in an upright standing position; the spine carries the bodyweight load, which affects the spinal posture. Upright standing radiographs are used for studying degenerative spine disorders and osteoporosis. The body is a 3D object, but radiographs are 2D. Thus, some information is lost. In spinal radiographs, the spine is overlaid with soft tissue and ribs; this makes detection and diagnosis more difficult. However, radiograph devices are faster and more accessible compared to other imaging apparatus.

2.1.2 Biplanar X-rays

EOS is a 2D/3D X-ray imaging system, which takes simultaneous anterior-posterior and lateral 2D entire body images. Since in this imaging system, two orthogonal view images are taken at the same time, the resulting images could be used in 3D reconstruction using statistical models, which has applications related to scoliosis and sagittal balance [1]. This imaging technique is validated for studying pelvic and lower-limb deformity, and pathology in adult and pediatric populations [1].

One of the main advantages of this technique is that the patient is exposed to a lower dose of radiation than conventional X-rays. Although this imaging offers the possibility to reconstruct 3D models from the images, a trained operator is required for this purpose, and it is not done automatically. The other drawback is that EOS technology is not cost-effective, hindering its applicability and thus the lack of its presence in clinical routine [2, 1].

2.1.3 Computed tomography

Contrary to radiographs, the CT imaging is 3D, but usually, the patient is in a prone position. Thus the spine is not carrying the body weight, and the curve is different from the patient's radiograph. However, we have the 3D information of the spine without occlusion. In a CT scanner, there is a rotating X-ray source combined with detectors. The emitted X-rays from the source traverse the object and then reach the detector. The X-ray attenuation depends on the materials in an object. While capturing the scan, the source and detectors are rotated, and all directions of the object are acquired. The difference in X-ray attenuation while traversing objects allows for an image reconstructing that reflects the absorption coefficients of different tissues by inverting the Radon transform [3]. Normally Hounsfield units (HU) of -1000 belongs to air, 0 HU represents water and bone is above 200 HU.

2.2 Clinical conditions

Throughout this dissertation, we introduce methods with applications related to the spine and cranial implant design. We review the relevant clinical conditions briefly in this section.

2.2.1 Back pain

The spine underpins the body and the organ design, and it has a significant part in movement and body load bearing. It plays a vital role in protecting the spinal cord from injury, and external mechanical stress [4]. Degenerative spinal disorders are a significant cause of back pain and are among the most common indications for spinal surgery. Prevalence of chronic back pain is one of the major causes of disability. Biomechanical factors mainly cause back pain [5, 4]. In aged patients, osteoporosis makes the diagnosis for the pain cause more complicated. Surgery is often required to treat instability-related pain and to restore the balance of the spine. However, when and how to perform surgery remains a highly subjective decision based on the surgeon's experience. Thus, designing methods for quantitative patient-specific biomechanical models is dearly needed. Studies on spine biomechanics are involved in different areas [4], quantitative imaging [6], Finite Element Modeling (FEM) of vertebrae [7], alignment analysis [8] and biomechanical models [9].

2.2.2 Cranioplasty

Cranioplasty is a surgery to repair injured skulls, where a part of the skull bone is removed. The damages to the skull could be caused by brain tumor surgery, severe head trauma, or bone loss due to infection. Fig. 2.1 illustrates of a defective 3D model of skull. A Patient Specific Implant (PSI) is required to re-establish the skull with desired mechanical and anatomical functionality and provide the patient with a higher quality of life in cranioplasty surgery [10, 11]. Designing PSI is time-consuming, and the surgery

is done after the implant is designed and produced. Thus, devising automatic and reliable implant design solutions will accelerate the surgery procedure and decrease the costs substantially [12]. Regarding the recent achievements in 3D printing, the designed 3D models of implants could be printed rapidly. Currently, for designing the cranial implants, the CT scan of the patient's head is used to segment the skull bone, and the segmentation is converted to a 3D model. Next, the resulting 3D skull model is utilized as a guide for Computer-aided Design (CAD) procedure of the implant [13]. The current approaches for designing the implants are based on the assumption of symmetry in the skull, while this assumption is not correct when the skull is deformed [10]. According to the recent advancements in shape processing using deep learning, there is a potential for employing these methods to process personalized implant design.

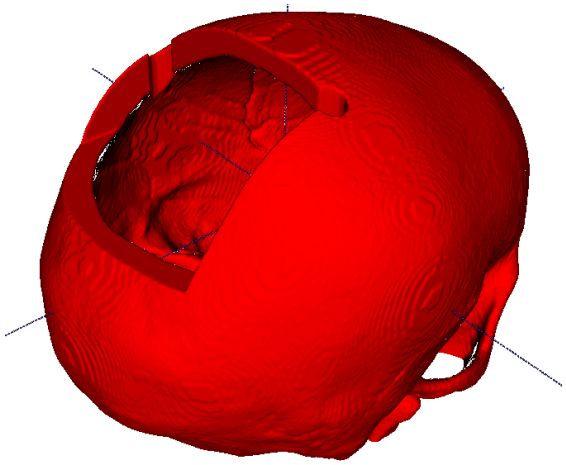


Figure 2.1: A 3D model of a defective skull.

2.3 Image registration

There are different tasks for processing and analyzing the medical images, including segmentation, registration, reconstruction. The focus of this dissertation is on registration and 3D reconstruction of bone shape. We introduce the tasks and review the related work in this section. Image registration or image matching is the process of aligning two or more moving images to a target image. The purpose of this process is to transform the input images that best align the target image. Image registration has important applications in medical image analysis for image fusion or segmentation. Image fusion using registration includes aligning images from the different modalities (like CT, MRI and Positron Emission Tomography (PET)), different time stamps. Thus, integrating the useful information from two or more images. Image registration is also used for anatomical segmentation in medical imaging by deforming shape templates, namely atlas, to a target image and generating the corresponding segmentation map. Image registration

has been applied for different human organs in 2D and 3D images: brain, retina, spine, heart, spine, pelvis, vascular structures, bones, knee, prostate, lung [14].

In image registration, the term fixed image or target image is defined as the image the remains unchanged, and the moving image is defined as the image which is transformed to match the fixed(target) image. A significant number of registration algorithms have been proposed. However, in any registration procedure, image modality selection, the feature space, a similarity metric, a transformation function, and an optimization algorithm are common steps. Registration algorithms could be classified into different categories depending on the data dimensionality, transformation elasticity (rigid, affine, deformable), learning-based models vs. optimization-based algorithms. Fu et al. [15], and Oliveira et al. [14] did comprehensive reviews on registration solutions and provide taxonomies of different deep learning-based registration approaches and traditional approaches for registration.

Registration algorithm could be classified based on the feature space information. The feature space could be the raw voxels, the intensity gradients, or structural information extracted from images, like segmentation maps, edges, surfaces, or graphs. Some algorithms work on the information from the frequency transformation of images. Registration algorithms could also be labeled as global or local registration if they leverage the entire image or focus on parts of the image.

In optimization-based registration methods, an optimization algorithm is employed to search iteratively to find the optimal geometric transformation to warp the moving image and match the target image. Based on the similarity metric or the objective function (cost function), the optimization algorithm could be defined as a minimization or maximization task. Usually, an initial alignment or pre-registration leads to a faster converges of the registration.

The iterative searching nature of optimization-based registration algorithms makes it time intensive. Some methods are proposed for leveraging GPU computing power [16] to accelerate this process.

For minimizing the objective function, either the similarity measure between the images is used, or some defined landmarks are extracted. The distances between the corresponding landmarks in the target and moving image are subjected to minimize. The detected landmarks and extracted features are required to be invariant to geometrical transformations to set up a reliable registration pipeline.

The other criterion for classifying the registration algorithms is based on the geometric transformations applied to moving images. The geometric transformations are divided into rigid and deformable categories. In rigid transformation we have 6 Degrees of Freedom (DOF) for a 3D space. Three of the parameters are for translation and 3 for rotation. The non-rigid transformation class includes the similarity transformation (translation, rotation, and uniform scaling), affine (translation, rotation, scaling, and shear), projective, and curved. The curved transformation is also called a deformable transformation.

The affine geometric transformation is usually defined with homogeneous coordinates. Thus, a $4x4$ matrix is used to represent the transformation. In the following, we define

an affine transformation including scaling and rotation about each axis:

$$T_{affine} = R(\theta_x) * R(\theta_y) * R(\theta_z) * T(S), \tag{2.1}$$

$$R(\theta_x) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) & 0 \\ 0 & \sin(\theta_x) & \cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2}$$

$$R(\theta_x) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) & 0 \\ 0 & \sin(\theta_x) & \cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.3}$$

$$R(\theta_y) = \begin{bmatrix} \cos(\theta_y) & 0 & \sin(\theta_y) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta_y) & 0 & \cos(\theta_y) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.4}$$

$$R(\theta_z) = \begin{bmatrix} \cos(\theta_z) & -\sin(\theta_z) & 0 & 0 \\ \sin(\theta_z) & \cos(\theta_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

$$T(S) = \begin{bmatrix} S_x & 0 & 0 & 0 \\ 0 & S_y & 0 & 0 \\ 0 & 0 & S_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.6}$$

$$Affine_{vf} = \tau(T_{affine}), \tag{2.7}$$

where $R(\theta_x)$, $R(\theta_y)$, $R(\theta_z)$ represent rotation about X, Y, Z axes and and $T(S_x)$, $T(S_y)$ and $T(S_z)$ represent the scaling factor for each dimension. The $Affine_{vf}$ is the affine vector field, and $\tau()$ is the function to generate an affine vector field given the transformation matrix.

Rigid registration is usually used to match rigid objects like bones or applied as a pre-registration step before a deformable registration. The latter use case's importance is to limit the search space and avoid converging to a local minimum before running a complex registration with more number of parameters to be optimized [17, 18, 14].

Similarly, the affine registration is also used as a pre-registration step. In some cases, it is applied as the second step registration after rigid transformation and before deformable registration [19, 14].

These pre-registration steps are increasingly necessary when dealing with low-resolution data with a low signal-to-noise ratio, like ultrasound or X-ray images. They help apply a rough alignment of moving image and target image before applying the complicated transformation with higher degrees-of-freedom. Especially in medical imaging, where

most human anatomy are deformable structures, reliable non-rigid registration algorithms are of high importance.

Deformable registration could be defined as free-form transformation. Any deformation is allowed, and controlled deformations, where the physical properties could be considered constraints while calculating the transformation parameters. Free-form algorithms are ill-conditioned and hard to regularize. Usually, in free-form deformation algorithms, a grid of control points is defined to regularize the deformation. To determine the local deformations, the points in that grid are moved individually in a way that optimizes the objective function. The transformation between control points is calculated by interpolation. There are different options for interpolating the transformation for the points between the control points, e.g., linear interpolation, nearest neighbor, B-spline interpolation [20, 21, 22, 23].

In B-spline registration, the image transformation is defined as a combination of basis functions. During the registration process, the optimal coefficients for these basis functions maximize the similarity measure.

Some deformable registration approaches model the deformation as an elastic transformation of a set of solid objects. Thus the entire group is deformable while it is solid locally. One example of such an item is the spine, a deformable object but composed of vertebrae, which are rigid objects. The idea is that the internal elastic forces of the solid oppose the deformation, and the external forces driven by the similarity measure try to deform the data to fit the target shape. The other deformable registration approach based on solid properties of object is Thin-plate splines (TPS) methodology [24, 25]. In TPS based registration the control point movements are constrained to bending energy [26].

An essential property of deformable registration algorithms is preserving the topology of structures in the images. In other words, the transformation is required to be diffeomorphic. It means that the transformation should be invertible and differentiable mapping with differentiable inverse [14]. Dalca et al. in [27] and Krebs et al. in [28] propose probabilistic deep neural networks for diffeomorphic registration. The free-form deformable registration algorithms can be diffeomorphic if regularized by a penalty term to the similarity function to avoid non-smooth deformations.

2.3.1 Similarity measures

Generally, the objective function for optimization in deformable registration problems is composed of two terms. The first term describes the voxel level similarity or structure similarity. While the second term is used as a regularization term and aimed for penalizing and smoothing the deformation field [23, 14, 15].

Deep Learning has transformed medical image analysis recently, with the state of the art performance in medical image segmentation, classification, registration, and reconstruction. Given large-scale datasets, deep neural networks can automatically extract intermediate, and high-level features from image data automatically [29]. Automatic feature extraction leads to superior performance compared to the machine learning methods working with handcrafted features. Generally, the building blocks of a registration work-

flow are an image deformation model, an objective function for measuring the agreement or alignment of images, and an optimization algorithm for minimizing the objective function. For training a deep neural network to register a moving image to a target image, in a supervised fashion, we need a dataset of aligned images. However, recently there were works on training the networks in an unsupervised scheme. For this goal, they employ the conventional intensity-based registration metrics into a learning problem to update network parameters for registration. The popular metrics for measuring the image alignments are normalized cross-correlation [30], mean-squared error [27] and LCC metric [28]. Currently, the most important limitation of registration algorithms is the traditional image agreement functions [31].

Thus, studies on learning metrics for this task are of interest. There have been studies on deep metric learning for registration with both supervised and unsupervised approaches [32, 33, 34]. One strategy for this purpose is to train the network to distinguish the registered and unregistered images. However still, one major limitation of these methods is the requirement for aligned training data. This requirement is increasingly necessary when working with multimodal images that are not obtainable simultaneously.

Spatial Transformer Networks introduced in [35], are neural networks for predicting the transformation given image data. Once the network predicts the spatial transformation parameters, they employ a differentiable image resampling method to apply the transformation. Next, based on the application, the resampled image is fed to another sub-network for a down-stream task. In [23] for example, they employed the Spatial Transformer Networks (STN) idea for segmentation medical images by registering a shape template. First, a convolutional network is used to estimate the transformation parameters. Next, the shape template is transformed by differentiable resampling, and the resulting shape is compared to the target segmentation mask. Finally, the error is backpropagated to the previous layers to update the parameters.

One of the most common tasks in medical imaging is the automatic alignment of two or more medical images from the same modality or different imaging modalities. Alignment is usually done by defining a metric for measuring the alignment between the pair of images and an optimization algorithm for maximizing the alignment or minimizing the disagreement between the moving and target image [15].

2.3.2 Atlas and registration based segmentation

Segmentation by registration is a popular method in the medical imaging community. There are various algorithms introduced for this purpose[23]. One common approach is multi-atlas registration, in which a dataset of atlases with labels and a similarity metric is used to select matching images. While testing, images are compared to examples in the atlas dataset, and the labels of matching images are used as segmentation candidates. This approach is not precise enough; thus, a refinement step could be applied to the selected segmentation candidates. It could also be used on images at patch level [36]. In [37] and [38] authors propose template registration methods to segment medical images.

2.4 Shape modeling

Shape models and priors are popular in 3D computer vision, and medical imaging and are widely used for shape reconstruction. The shape models are mainly generated based on a population of shapes manually and not acquired by a learning-based solution. In recent years, there have been works to show that generative models like Variational Autoencoder (VAE)s and Generative Adversarial Network (GAN)s could be used to generate and learn 3D shapes with details [39, 40, 41].

2.4.1 Statistical shape model

Statistical Shape Model (SSM) is widely used for reconstructing anatomical shapes (e.g., bones) in medical applications. A statistical model bone represents the average shape and shape variations of a given shape population. The idea is to reduce the dimensionality of the shapes and model every data sample using a limited number of control points. For example, reconstructing a complete skull given a defective shape or partially visible shape is the task of finding the set of parameters. The resulting shape best matches the input shape, except in the defective region. One common approach is using Principle Component Analysis (PCA) and defining shapes as a linear combination of shape components.

Recently deep learning solutions have emerged in segmentation, shape completion, registration, and shape reconstruction. We review the state-related work in each task using deep learning briefly in the following.

One of the shape representation methods is the mesh-based format. Typically, mesh-based models are used for simulating complex physical systems for different applications. Pfaff et al. [42] propose MESHGRAPHNETS, a deep learning method for mesh-based simulation with graph networks. Their solution is a mixture of encoder and decoder networks with a message passing mechanism; their model is trained to pass messages on mesh graphs.

2.4.2 Shape completion

3D shape completion is one of the fundamental fields of study in computer vision and has critical medical imaging applications. 3D shape completion from sparse annotation is a special case of 3D shape reconstruction from a single-view 2D image [43]. A major number of papers on this topic focus on 3D reconstruction. 3D shape reconstruction from one or multiple 2D images is an ill-posed inverse problem [43]. One common approach in 3D computer vision is to integrate shape priors into models. However, recently with the advancements in deep learning, the models can learn the 3D shapes and generate new images, given large-scale datasets. Our focus is on generating 3D shapes given 2D images with sparse annotations and completing 3D forms given defective shapes throughout this thesis. The former topic aims to infer the 3D spine posture from 2D radiographs, and the latter topic aims to design cranial implants automatically.

The current solution for 3D shape completion is based on either data-driven or learning-based methods. In a data-driven approach, the shape prior is defined based on learned shapes, and the shape completion is formulated as an optimization problem over the latent space [44, 45, 46, 47]. In general, these approaches perform better on real-world data. On the other hand, learning-based solutions work with full direct supervision to learn the shape of synthetic data [48]. The learning-based approaches are faster as they can infer the 3D shape in a single forward pass, compared to the iterative optimization process of shape inference in data-driven approaches. But one of the drawbacks of learning-based methods is the requirement for full supervision and demanding large datasets.

Several deep learning-based solutions have been introduced for the problem of shape completion. In learning-based solutions, the shape completion is done by supervision from synthetic data. To tackle the memory limit problem, in [49] authors use octrees to work with higher-resolution shapes. Other approaches either work on low-resolution shapes in voxel space (for example, 32^3) or take a patch-based approach to work on high-resolution results. Working on real-world data, accessing full supervision is not feasible. To alleviate this problem, most of the proposed models are on synthetic data or leverage the shape priors to constrain the space of possible shapes [48].

In medical image, there exists prior work on inferring 3D segmentation from 2D annotations in [50] they do 3D segmentation using sparsely annotated volumes, and [51] segmentation based on Maximum Intensity Projection (MIP) images. However, they require the input image and do not put any constraint on the anatomical shape.

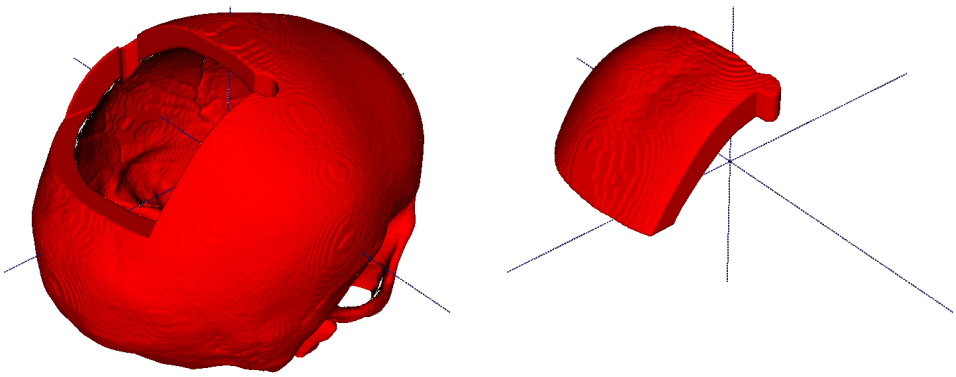
2.4.3 3D reconstruction from 2D images

In recent years, 3D reconstruction from 2D images has been a hot topic for computer vision researchers. In [41, 52, 53, 54] propose methods for 3D shape reconstruction from single view images working with full supervision. It means that the dataset required to train those models is made of images and corresponding 3D shapes in pairs. These models mostly work on synthetic data. In [55, 56, 57, 58, 59] they take advantage of self-supervision by enforcing reconstruction consistency from different views. In [48] the object types are used to employ a weak supervision training scheme.

2.4.4 Shape completion for cranial implant design

As introduced earlier in this chapter, automatic cranial implant design has been an under-researched field. Recently Li et al. [60] proposed an automatic methodology for cranial implant design. They also organized a challenge on automatic cranial implant design in conjunction with MICCAI 2020. As a result, some deep learning-based solutions were proposed for this problem. In this challenge, the problem statement is to design models to infer the 3D cranial implant shape given the defective skull as input. Fig. 2.2 presents a defective skull and corresponding cranial implant. The immediate solution is to find a method to complete the defective skull shape first. Next, the difference between the predicted skull shape and the defective skull represents the cranial implant. To this

end, the proposed methods were mostly based on encoder-decoder architectures with skip connections, similar to U-net [61]. For example, in [62], adopted a coarse-to-fine framework to generate the target implants in two steps; First, they employ a deep neural network to predict the coarse shape of the implant in 3D. Next, another network is trained for refining the predictions from the first step in 2D.



(a) defective skull **(b)** cranial implant

Figure 2.2: A 3D model of defective skull and corresponding cranial implant.

2.4.5 3D reconstruction of spine from 2D images

The balance of the spine is an essential factor for the development of spinal degeneration, pain, and the outcome of spinal surgery. It must be analyzed in an upright, standing position to ensure physiological loading conditions and visualize load-dependent deformations. Despite the complex 3D shape of the spine, this analysis is currently performed using 2D radiographs, as all frequently used 3D imaging techniques require the patient to be scanned in a prone position.

Biomechanical load analysis of the spine in an upright standing position is highly warranted in various spine disorders to understand their cause and guide therapy [63]. Typical approaches for load estimation either use a computational shape model of the spine for all patients or obtain a subject-specific spine model from a 3D imaging modality such as MRI or CT [64]. Although MRI and CT images can capture 3D anatomical information, they need the patient to be in a *prone* or *supine* position (lying flat on a table) during imaging. However, to analyze the spinal alignment in a physiologically upright standing position under weight-bearing, orthogonal 2D plain radiographs are the *de facto* choice. A combination of both these worlds is of clinical interest to fully assess the true bio-mechanical situation, i.e., to capture the patient-specific complex pathological spinal arrangement in a standing position with full 3D information [65, 64, 66].

As much spatial information is lost when projecting a 3D object in only two 2D planes, a random object cannot reliably be reconstructed from two orthogonal projections. However, the spine follows strong anatomical rules that are repeated only with

slight variations in any patient. Typical projections, i.e., lateral and a.p. radiographs, cover most of these variations, both on a local (per vertebra) and global (overall spinal alignment) level.

Literature offers a wealth of pre-existing registration-based methods for relating 2D radiographs with 3D CT or MR images. In [66], the authors propose a rough manual registration of 3D data to 2D sagittal radiographs for the lumbar vertebrae. For the same purpose, in [67], manual annotations of the vertebral bodies are used as a guideline for measuring the vertebral orientations in the upright standing position. These methods are time and manual labor-intensive and thus prone to error. Moreover, both of these works use only the sagittal radiographs for vertebra positioning while ignoring the coronal reformation, a strong indicator of the spine's natural curvature. Aiming at this objective, [68] introduced an automatic 3D–2D spine registration algorithm. The authors propose a multi-stage optimization-based registration method by introducing a metric for comparing a CT projection with a radiograph. However, this metric is hand-crafted, parameter-heavy, and not learning-based, limiting its generalizability and inference speed. In [69, 70, 71], the 3D shape of the spine is reconstructed using a biplanar X-ray device called 'EOS'. The system's advantage is the low radiation dose required and that both projections are acquired simultaneously, allowing for a direct spatial correspondence between the two planes. Hindering its applicability is the high device cost and thus the lack of its presence in clinical routine. Recently, the problem of reconstructing 3D shapes given 2D images has been explored using deep learning approaches. An approach was proposed in [72], where they introduce a deep neural network to synthesize 3D CT images given orthogonal radiographs using adversarial networks. However, this model is highly memory intensive and fails to synthesis smaller anatomies like vertebrae in 3D. Moreover, it has been evaluated only on DRR, and its clinical applicability in real radiographs remains to be validated. In [73] and [74] the authors design a model to generate a 3D given multiple arbitrary view 2D images. But the input images include only the object of interest without background, which does not apply to medical images like spinal radiographs. In our previous work for inferring the 3D standing spine posture [2] we introduce a new deep neural network architecture to combine the 2D orthogonal image information and reconstruct a 3D shape. Since the vertebrae are heavily occluded by soft tissue and ribs in the radiograph, the reconstruction is challenging. In some cases, the output shape is far from a vertebral shape.

The idea of incorporating shape priors in the model have been explored by [23] [75] [76] [77]. However, considering only the vertebral shape priors for reconstructing spine posture is not enough for our application. Most of the ideas proposed for using shape priors in deep neural networks are for a single object. In contrast, the spine shape is more complicated as it is made of multiple objects (vertebrae) connected and deforming subject to human anatomy constraints. Thus, defining a prior for explaining intervertebral constraints and spinal shape is crucial here.

A few works have modeled the spinal shape. For instance, in [78] authors designed an automatic framework segmenting vertebrae from arbitrary CT images with a full spine surface model. To create the model, they first scanned a commercially available plastic phantom to create the template, and then they manually registered it to ten

actual scans of the entire spine. In [79] authors 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 [80], graph models [81], hidden markov models [82] and hierarchical models[83][84][85] have also been proposed. For example, Ruiz et al. [80] 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. [81] 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. But non of the works mentioned above are incorporated in a deep learning model for training and inference.

2.5 Spine image analysis tools

Biomechanical analysis of the spine requires patient-specific 3D models and biomechanical representation of vertebral joints. The current approaches for this analysis are mostly done manually, which makes it highly time-consuming. Thus devising methods to process these data automatically is beneficial. [86] is one of the first works published recently to automate the process of generating patient-specific biomechanical models of the spine for surgery.

We developed a set of scripts for processing the spinal data. This set of tools were used in the works introduced in the following chapters of this dissertation. Thus, introducing the data processing tools would ease understanding the workflow in the following chapters. The tools introduced here have also been used for other works on spine biomechanical analysis that will be published in the future.

Here we introduce the labels defined for each vertebra and vertebral subregion. Table 2.1 gives an overview of the vertebral labels and Table 2.2 presents the labels used for defining vertebral subregion.

Vertebral levels	label values
C1-C7	1-7
T1-T12	8-19
L1-L6	20-25

Table 2.1: Vertebral labels.

2.5.1 Vertebral subregion segmentation

For segmenting the vertebral subregions, we took an atlas registration-based segmentation approach. We used the labels introduced in Table 2.2 for defining the vertebral

vertebral subregions	frontal direction	axial direction	lateral direction	label value
vertebral body complete	-	-	-	40
vertebral arch	-	-	-	41
spinous process	-	-	-	42
transverse process	-	-	left	43
transverse process	-	-	right	44
articulate process	-	superior	left	45
articulate process	-	superior	right	46
articulate process	-	inferior	left	47
articulate process	-	inferior	right	48
vertebral body vertical cortex	-	-	-	49
vertebral body whole trabecular compartment	-	-	-	50
vertebral endplate	-	superior	-	52
vertebral endplate	-	inferior	-	53

Table 2.2: Vertebral subregions labels.

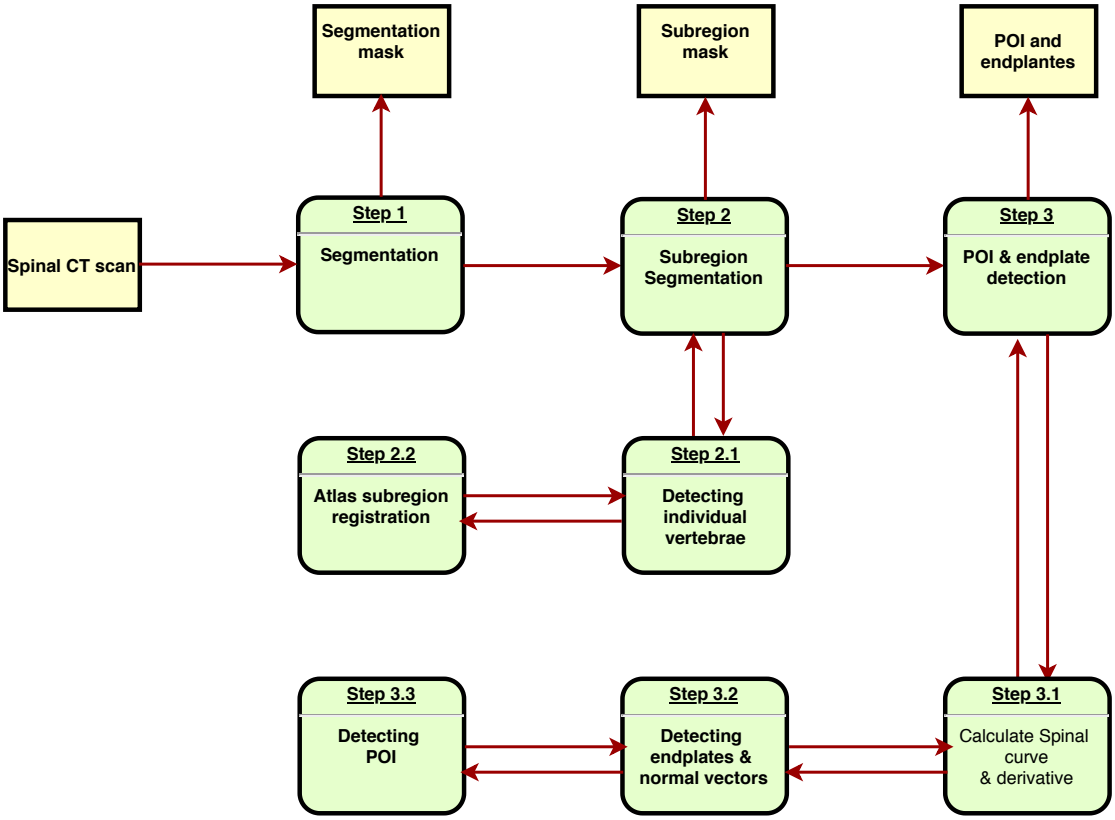


Figure 2.3: Vertebral subregion segmentation and muscle insertion points detection pipeline block-diagram. The yellow boxes are the input/output files, while the green boxes are processing steps.

subregions in the atlas. Given a new CT scan, we follow the procedure below to segment subregions of vertebrae automatically.

- First, we segment it automatically using a deep neural network trained on a large-scale vertebra segmentation dataset [87]. In the resulting segmentation map, each vertebra has a distinct integer intensity, indicating its label as explained in Table 2.1.
- We detect each vertebra as a separate object and extract a bounding box around it.
- For each vertebra, we select the corresponding vertebra from the atlas and register it to the target with 6 DOF, namely "rigid registration". We optimize the alignment between the atlas vertebra and target vertebra by updating the location in 3D (X, Y, Z) and Rotation about each axis. Once the registration is done, we apply the transformation to the atlas subregions to match the transformed atlas mask.

- Aiming to increase the match between the transformed atlas and the target vertebra, we take the transformed mask from the previous step and apply an affine registration 12 DOF to match the target vertebra. Once the transformation is calculated, we apply it to the corresponding vertebral shape with subregions.

- We take a step further and apply to a deformable registration on the previous step results to improve the fine shape details, wherein the number of parameters to optimize is equal to the number of voxels in the image times three.

Fig. 2.3 illustrates the steps for subregion segmentation procedure.In the following, we elaborate on each step introduced above. Fig. 2.4 depicts an example of spine segmentation map and Fig. 2.5 shows the corresponding vertebral subregion segmentation image.

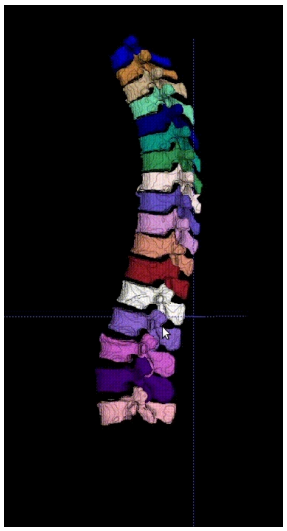


Figure 2.4: A spine segmentation example.

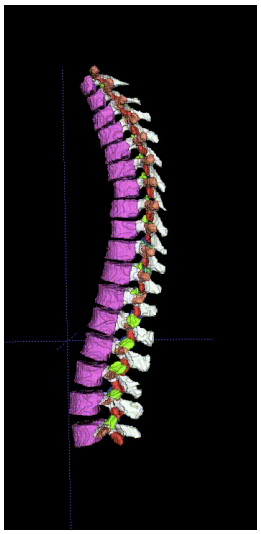


Figure 2.5: Vertebral subregion segmentation of the spine.

2.5.2 Optimal spinal image viewer

Evaluating 3D scans of the spine and segmentation quality is a cumbersome task. Summarizing the essential 3D shape information and visualizing the result in a 2D snapshot can accelerate this procedure significantly. This idea particularly helps when creating a high-quality, large-scale dataset for training Deep Learning (DL) models. Due to different vertebral orientations, selecting the mid-slice of a scan and the segmentation image would not be enough for visualizing all vertebrae completely. In Fig. 2.6 this problem is demonstrated. To alleviate this issue, we define a curved 3D surface to cover most of each vertebra and unfold it on a 2D plane. The steps for this task are as follows:

- Segmenting the subregions as explained in the last subsection.

- Calculating the center of mass of the vertebral body and vertebral process for each vertebra.

- Fit a spline to the sets of vertebral body center of mass and vertebral center of mass

- Calculating the line connecting the corresponding voxels on the curves calculated earlier.

- the set of lines in the preceding step, represents a curved plane in 3D, sampling the corresponding voxel in image and segmentation and projecting to 2D yields the snapshot with the maximum possible view of each vertebra.

Fig. 2.7 shows some examples of generated snapshots.

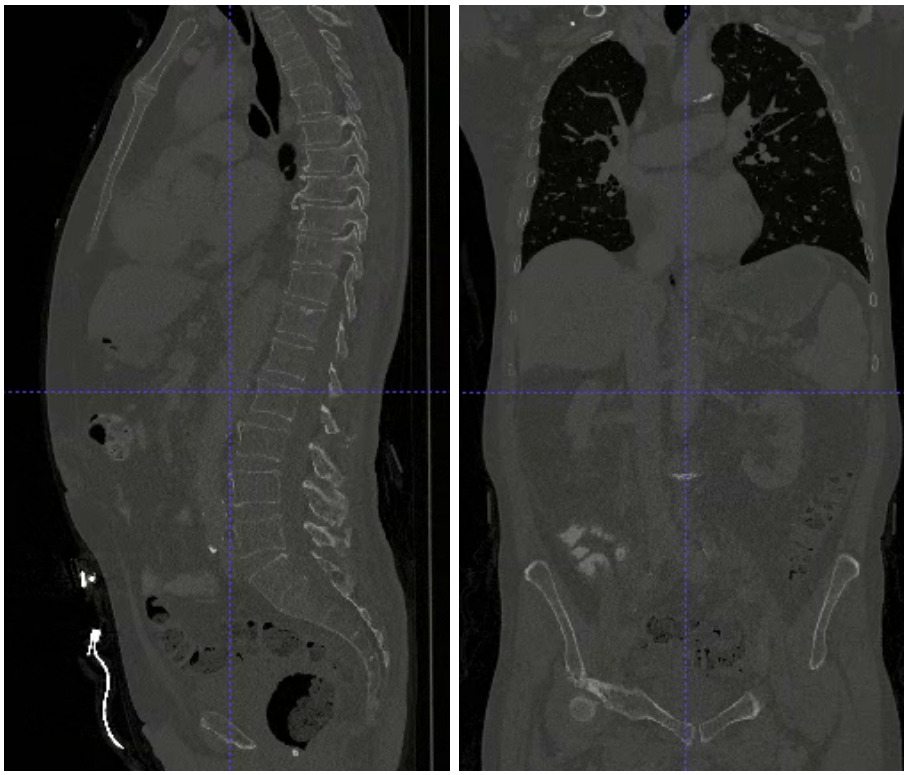


Figure 2.6: Examples spinal of CT slices from different views.

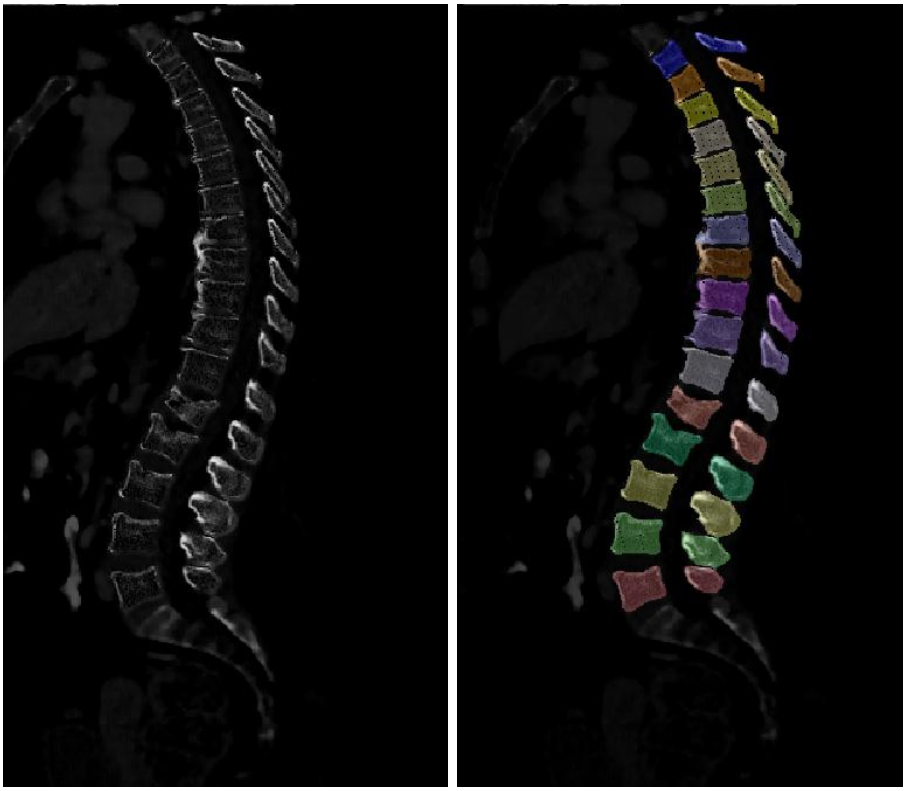


Figure 2.7: Snapshots with complete spine visibility. We define a curved 3D surface to cover most of each vertebra and unfold it on a 2D plane. The left image depicts the raw image, and the right one is the raw image overlaid with segmentation maps.

2.5.3 Vertebral endplate detection

Given the vertebral subregions, we separate the vertebral bodies and detect vertebral endplates by making projections along the spinal axis. Detecting endplates automatically facilitates the procedure for biomechanical analysis.

2.5.4 Muscle insertion point detection

After detecting the subregions, we detect the muscle insertion points by selecting the outermost points of subregions. These points are also required for biomechanical analysis. In Fig. 2.8 the detected muscle insertion points for all vertebra are visualized.

Detection of the vertebral endplates and muscle insertion points is a starting point for individualized therapy planning in back pain patients. Individualizing biomechanical models allows personalized therapy planning and outcome prediction after spine surgery.

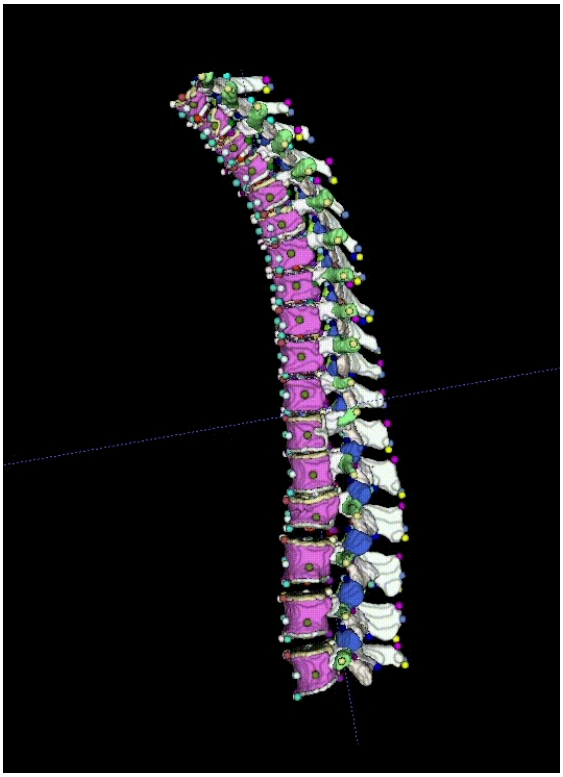


Figure 2.8: Vertebral landmarks and subregions. The subregions and labels are explained in Table 2.2.

2.5.5 Output for multibody simulation

Simpack [88] is a Multibody System Simulation (MBS) software designed for simulating the non-linear motion of the mechanical systems. The dynamic motion, coupling forces, and stress in 3D models could be predicted and visualized using this software package. Simpack is mainly applied in the automotive, engine railway, and wind energy industry sectors. However, this software could be used for other applications, including biomechanical simulations. We developed a script to convert the spine analysis toolbox's output to a Simpack readable input. After that, the generated input data could be used to start the simulation in Simpack. Thus, the entire process from imaging to biomechanical analysis could be done automatically for each patient.

2.5.6 Cobb angle calculation

For evaluating and treating scoliosis, the spinal curve is required. Cobb angle is a measure of the spinal deformity in scoliosis. Cobb angles estimation is currently done manually given the spinal radiographs, as demonstrated in Fig. 2.9. This process is labor-intensive and prone to error. There are studies to automate this process on spinal radiographs [89]. Using the spinal landmark detection procedure introduced in the pre-

Figure 2.9: AP view spinal radiographs are used for scoliosis evaluation.

vious section, we can estimate the spinal curve and then calculate the Cobb angles automatically in 3D data. To calculate the Cobb angles, first, we estimate the spinal curve by fitting a curve to the vertebral centroids. Next, we calculate the normal to the curve at each vertebral endplate. The angles between the normal vectors at certain vertebrae give us the Cobb angles. Cobb angle is a measurement of the degree of side-to-side spinal deformities in scoliosis.

Once the centroid of vertebral bodies is calculated, we follow the steps below to calculate the Cobb angles.

- In the first step, we fit a curve of degree 3 to the centroids of vertebrae and calculate the first derivative of the curve.

- Next, we calculate the maximum and minimum slope of the curve with respect to Z dimension.

- We calculate the normal vectors to the curve at the points detected from the previous step.
- The angle between the vectors from the previous step determines the Cobb angle.

2.5.7 Digitally reconstructed radiographs

The size of objects (towards the scan's periphery) is larger in radiographs compared to their true size due to the cone-beam of the gamma-ray source. Fig. 2.10 depicts this effect. Generation of the DRR is performed using a ray-casting approach, wherein a line is drawn from the radiation source (focal point) to every single pixel on the DRR image, and the integral of the CT intensities over this line are calculated.

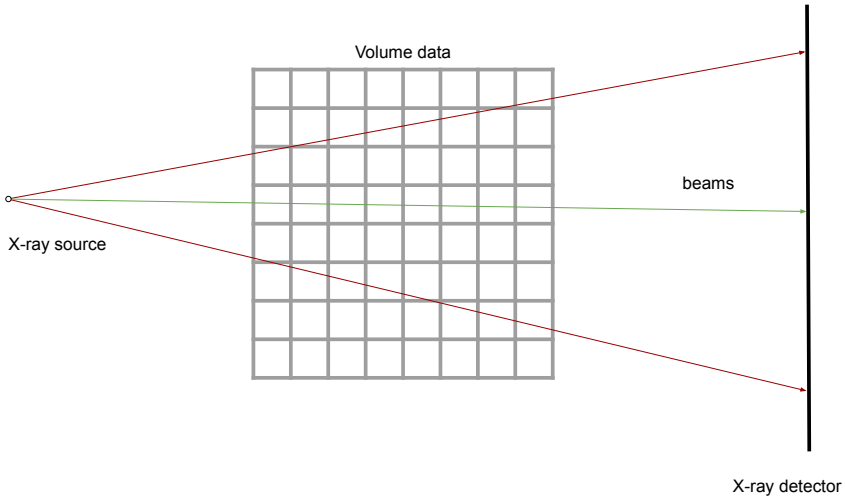


Figure 2.10: Schematic X-ray imaging. The size of the object is larger in radiographs compared to their true size due to the cone-beam of the gamma-ray source.

For training the model introduced in Chapter 3 of this dissertation, we needed to train the model on synthetic data. We simulated the spinal radiographs from CT scans. Parameters for this generation include the radiation source-to-detector (= 180cm in this work) and the source-to-object distance (= 150cm here). In Fig. 2.11 a sagittal and a coronal view DRR generated from a CT scan is illustrated.

Figure 2.11: Sagittal and coronal view DRRs, generated from a 3D CT image.

3 Vertebral Labelling in Radiographs: Learning a Coordinate Corrector to Enforce Spinal Shape

This chapter has been published as peer-reviewed paper and is presented here with minor modifications.

Bayat, Amirhossein, Anjany Sekuboyina, Felix Hofmann, Malek El Husseini, Jan S. Kirschke, and Bjoern H. Menze. "Vertebral labelling in radiographs: learning a coordinate corrector to enforce spinal shape." In International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging, pp. 39-46. Springer, Cham, 2019. DOI: 10.1007/978-3-030-39752-4_4

Synopsis: This work introduces a deep learning method for vertebrae localization and labeling in spinal radiographs. Detecting and labeling vertebra in the radiograph is challenging due to the large field of view, heavy tissue overlay, and noise. Most of the works for vertebra labeling were on CT images; our work was one of the first methods proposed for radiographs. We proposed a new architecture robust to occlusion by enforcing the spinal shape. In our approach, the model has a holistic view of the input image irrespective of its size. Our model predicts the labels and vertebrae locations in two steps: Firstly, a is used to estimate the vertebrae location and label by predicting 2D Gaussians. Then, we introduce the Residual Corrector (RC) component, which extracts each vertebral centroid's coordinates from the 2D Gaussians and corrects the location and label estimations by considering the entire image. The functionality of the RC component is differentiable. Thus, it can be merged to the deep neural network and trained end-to-end with other sub-networks.

contributions of thesis author: algorithm design and implementation, experiment design and composition of manuscript.

3.1 Abstract

Localizing and labeling vertebrae in spinal radiographs has important applications in spinal shape analysis in scoliosis and degenerative disorders. However, due to tissue overlaying and size of spinal radiographs, vertebrae localization and labeling are challenging and complicated. To address this, we propose a robust approach for landmark detection in large and noisy images and apply it on spinal radiographs. In this approach, the model has a holistic view of the input image irrespective to its size. Our model predicts the labels and locations of vertebrae in two steps: Firstly, a FCN is used to estimate the vertebrae location and label, by predicting 2D Gaussians. Then, we introduce the Residual Corrector (RC) component, that extracts the coordinates of each vertebral centroid from the 2D Gaussians, and correct the location and label estimations by taking into account the entire image. The functionality of the RC component is differentiable. Thus, it can be merged to the deep neural network, and trained end-to-end with other sub-networks. We achieve identification rates of 85.32% and 52.28% for sagittal and coronal views and localization distance of 4.57 mm and 5.33 mm in sagittal and coronal views radiographs, respectively.

3.2 Introduction

Localization and labeling in spine is applicable in Cobb angle calculation, surgery planning, bio-mechanical load analysis, diagnosing vertebral fractures or other pathologies. Analysing the shape of the spine in scoliosis and degenerative disorders should be performed in an upright standing position, which is captured using 2D radiographs. Thus, vertebrae localization and labeling is of crucial importance on radiographs, while most of the work for vertebrae labeling has been performed on CT scans.

Vertebrae localization and labeling in radiographs are challenging tasks, due to overlaying tissue and noise. Particularly, in lumbar spine the vertebrae are superimposed by heterogenious soft tissue. Thus the vertebrae's intensity looks different from the ones in the other regions. Another challenge is the large size of the spinal radiographs, as they include the entire spine. The latter issue makes it difficult to design a network with a receptive field, covering the entire image. For the same reason, applying available deep neural network architectures to process these images is sub-optimal. On the other hand, applying fully connected layers to address this problem is not feasible as well, as it leads to increasing the parameters, and limits the performance of the network only to fixed size images.

The related works for vertebrae localization and labeling are mostly on CT scans. Usually, an FCN is employed to estimate a heat map of points of interest, then, a second module is applied on the heat map to predict the locations and labels more accurately. The main difference between various methods is mostly in the the second module.

In [90] and [91] Yan et al propose a deep FCN being trained with deep supervision followed by message passing or convolutional LSTM to improve the predictions. In [92] the authors propose to regress a heat map and then regress the landmarks using FCNs.

In their approach, the field of view in the input images is variable. While, we always have complete spine image in the input. In [93] and [94] used regression forests label the vertebra, but the limitation of their model is limited field of view. To address the problem of limitation in field of view, the authors in [95] used Multilayer Perceptron (MLP) networks to capture the long-range context features. In the same direction, in [96] the authors introduce an adversarial framework for localization and identification of vertebrae in CT scans. They use an energy-based discriminator in an adversarial setting to correct the labels predicted by the FCN.

While the related works mentioned above, are on CT, we address the problem of vertebrae labeling and localization in spinal radiographs. This task is more challenging due to tissue overlay and noise. We devise a model with unlimited receptive field and enforcing the spinal shape. Our solution is a two-level supervision approach. First, we estimate the rough location of landmarks by predicting 2D Gaussians using a FCN. At this level, the 2D Gaussian image is compared to the ground truth 2D Gaussians. Next, we introduce the RC component, to convert the Gaussians predicted in the previous step, to coordinate format and then correct the coordinates. In this way, the dimensionality of the data decreases from $h \times w$ to 2×24, where h and w are height and width of the input image respectively and 24 is the number of vertebrae. For each vertebra, 2 parameters are estimated as 2D coordinates. In other words, we can decrease the dimensionality of data from input images with variable size to a fixed size of 2×24, as long as the input image includes the entire spine. The functionality of the RC component is differentiable. Therefor, the entire network is trainable in an end-to-end fashion.

Our model design, leverages both the texture features and spinal shape (inter vertebral spatial relation) to localize and label the cervical, thoracic and lumbar vertebrae. We train and test our model on two datasets. One including radiographs of sagittal view and the other one a dataset of coronal view radiograph.

3.2.0.1 Contributions.

1) We design a robust and accurate deep neural network architecture, with two level of supervision, applicable to landmark detection, on large images. 2) We design residual corrector (RC) component, a differentiable module to convert centroids predicted as 2D Gaussian images to coordinates, and then correct them. 3) Finally, we achieve identification rate of 80.55% for both coronal and sagittal view, localization distance of 7.71 mm and 7.59 mm for sagittal view and coronal view respectively.

3.3 Methodology

3.3.0.1 Annotations.

The output of the FCN is compared to ground-truth Gaussian images. Similar to [96], the ground-truth Gaussian image $Y \in \mathbb{R}^{(h \times w \times 25)}$ is a 25-channeled, 2D image with each channel corresponding to each of the 24 vertebrae (C1 to L5), and one for the background; h and w are height and width of the input image respectively. Each channel is constructed

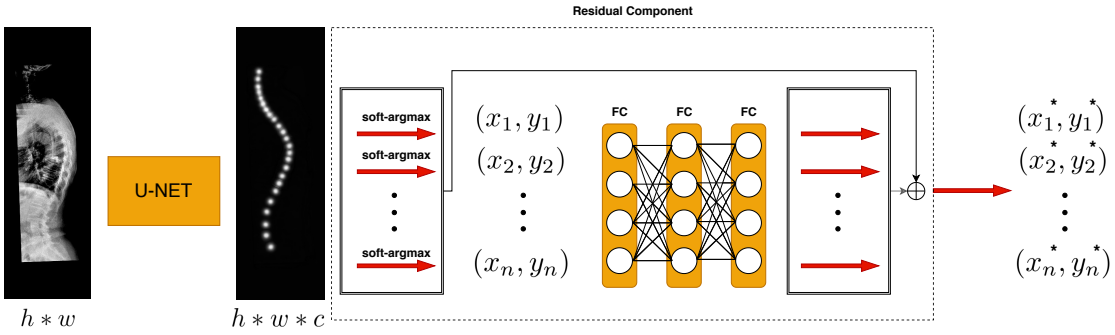


Figure 3.1: Overview of the model.

as a Gaussian heat map of the form $y_i = e^{-||x-\mu_i||^2/2\sigma^2}$, $x \in \mathbb{R}^2$ where μ_i is the location of the i^{th} vertebra and σ controls the spread. The background channel is constructed as, $y_0 = 1 - \max(y_i)$.

3.3.0.2 Architecture.

Fig. 3.1 gives an overview of our proposed model. It is composed of two sub-networks: U-net[61] as a FCN, and the RC component, which is highlighted with a dashed-line box in Fig. 3.1.

In the RC component, first the centroid coordinates are extracted using soft-argmax method. In this way, the dimensionality of the data decreases from $h \times w$ to 2×24, where 24 is the number of vertebrae. For each vertebra, 2 parameters are estimated as 2D coordinates. In other words, we can decrease the dimensionality of data from input images with variable size to a fixed size of 2×24, as long as the input image includes the entire spine. After that, we have a residual block of fully-connected layers to 1) correct the location of the estimated coordinates and 2) increase the receptive field of the model. Since the receptive field of the model covers the entire input image, it can capture long-range dependencies in the inputs with fewer network layers and parameters compared to the FCN approaches with similar receptive field. As soft-argmax is differentiable, we can train the fully-connected layers along with the FCN end-to-end and variable size input images. The fully connected layers in RC component sub-network, give a holistic view of the input image to the network and leads to global consistncy of the estimated vertebral coordinates and labels. i.e. the order of label and also morphological consistency of the spine.

3.3.0.3 Training.

The FCN sub-network, predicts 2D Gaussians, the predicted image in this level is compared to the ground truth Gaussian image. For this, we use the loss function introduced in [96], which measures the L_2 distance supported by a cross-entropy loss over the soft-max excitation of the FCN prediction and ground truth Gaussian image.

$$L_{Gaussian} = ||Y - \tilde{Y}||^2, \tag{3.1}$$

$$L_{ce} = H(softmax(Y), softmax(\tilde{Y})), \tag{3.2}$$

$$L_{img} = L_{Gaussian} + L_{ce}, \tag{3.3}$$

Where $\tilde{Y}$ is the predicted Gaussian image, Y is the target Gaussian image and H is the cross-entropy function. Next, using our RC component we extract the centroid coordinates and correct them. For this stage, we use a Smooth Absolute(Huber) loss to compare the predicted coordinates to the ground truth.

$$L_{coordinate} = \begin{cases} \frac{1}{2}||C - \tilde{C}||^2 & \text{for } |C - \tilde{C}| \leq 1, \\ \left(|C - \tilde{C}| - 1/2\right) & \text{otherwise.} \end{cases} \tag{3.4}$$

Finally, the total loss is the sum of the loss at the first and second stages.

$$L_{total} = L_{img} + L_{coordinates}, \tag{3.5}$$

3.3.0.4 Inference.

Once the model is trained, for inference the input image is fed to the FCN sub-network, the result is a multi-channel image with the same size of the input image. In the resulting image, each vertebral centroid is predicted as a 2D Gaussian in the corresponding channel. Next, this multi-channel image is fed to the RC component. In the RC component, the coordinates of the Gaussians are extracted using the soft-argmax function. Then, the extracted coordinates are corrected by the residual fully-connected layers.

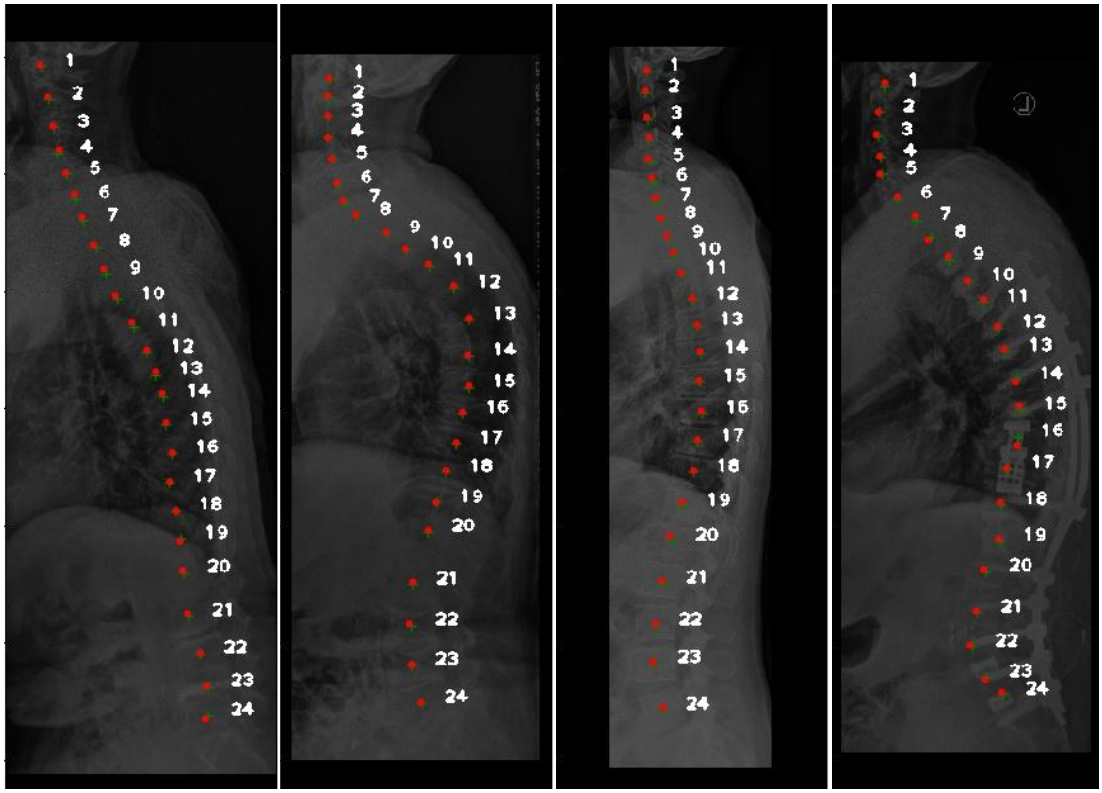
3.4 Experiments and Results

3.4.0.1 Datasets.

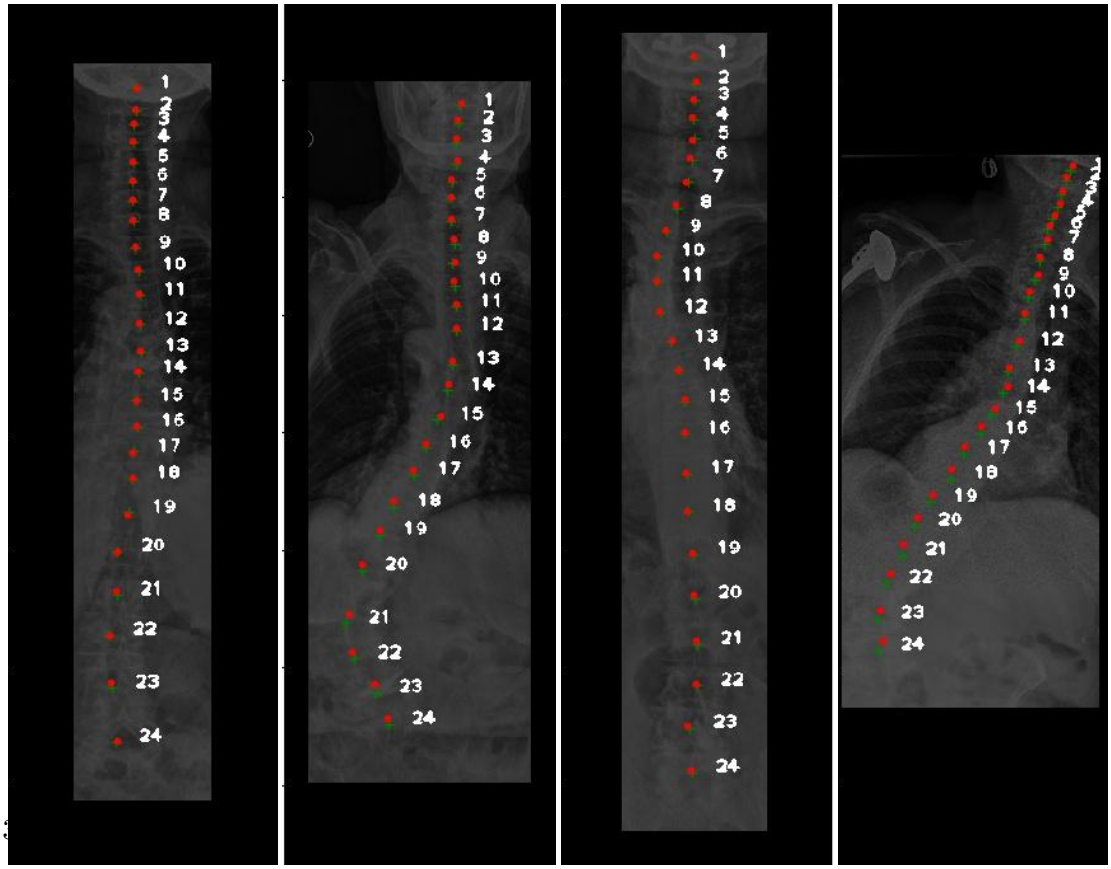
We work with an in-house dataset of 122 patients for each we have the coronal and sagittal radiographs, we also have the vertebral centroids annotations of them. We selected 100 cases randomly for training and 22 cases for validation. The radiographs are mostly from old patients, some cases with scoliosis or metal insertion. we augmented the data by rotation the images 2, -2, 5 and -5 degrees. We create two separate datasets out of this dataset, one for coronal view and the other one for sagittal view radiographs.

3.4.0.2 Experiments.

We carry out four experiments to validate our approach: 1) First, we train only U-net to localize and label the vertebral centroids on sagittal view. 2) We train U-net on coronal view radiographs. 3) Next, in order to improve the performance of the model, we add



(a) Sagittal view radiographs



(b) Coronal view radiographs

Figure 3.2: Testing the model on sagittal coronal radiographs. The predicted centroids are plotted in red dots and the ground truth centroids in green crosses, the numbers indicate the vertebral label.

the RC component to the model and train the model sagittal view. 4) We train the model with RC component on coronal view radiographs.

We implement our network in the Pytorch framework and use a Quadro P6000 GPU for training the model. In all experiments, the initial learning rate is 0.0001 and Adam is used as optimizer, and the models are trained for 150 iterations.

3.4.0.3 Evaluation.

To evaluate the performance of our network, we use the identification rates(id.rate) and localisation distances (d_{mean} and d_{std}) in mm [93]. Table 3.1 compares the performance of U-net and our model, trained on sagittal view radiographs. Similarly, Table 3.2 reports the performance of U-net and our model, trained on coronal view radiographs.

Measure	U-net	U-net+RC
Id.rate	61.3%	**85.32%**
d_{mean}	11.54 mm	**4.57** mm
d_{std}	12.73 mm	**3.84** mm

Table 3.1: Quantitative performance comparison of U-net and our model on sagittal view radiographs.

Measure	U-net	U-net+RC
Id.rate	44.5%	**52.63%**
d_{mean}	12.32 mm	**5.33** mm
d_{std}	11.12 mm	**5.94** mm

Table 3.2: Quantitative performance comparison of U-net and our model on coronal view radiographs

We did not compare our results to vertebrae labeling methods on CT, as they are tested on another modality. The results suggest that the residual correction approach, improves the performance of the model significantly. Due to the limited receptive field of U-net, the localization and labeling are not accurate. The RC component takes the entire spinal shape into account and learn the inter vertebral spatial relation. Therefore, it can correct the predictions of U-net, by enforcing the true shape of spine.

Fig. 3.2 demonstrates examples of testing our model on coronal and sagittal images. The results suggests that our model is robust against metal insertion and treatment effects visible in the radiographs. Also it is robust against shift and small rotations. As it is shown in Fig. 3.2, our model handles variable size radiographs in both sagittal and coronal view. Finally, our model performs on the raw image data, without preprocessing or employing another network to localize the spine and all of the calculations are done in a single forward pass.

3.5 Conclusion

Processing large medical images using fully convolutional networks is suboptimal due to limitation of the receptive field and large size of spinal radiographs. Also, it is not feasible to apply fully-connectd layers to these networks to compensate the limitation of receptive fields, as it increases the number of parameters, significantly. Downscaling the images leads to losing details and makes the network prone to error. We propose an architecture to extract the local features using a fully convolutional network and learn the global shape of the spine using a residual corrector module. Finally, our model is robust against treatment effects in radiographs and can localize and label the vertebra in a single forward pass.

3.5.0.1 Acknowledgements

This work was funded from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (GA637164–iBack–ERC-2014-STG).

4 Inferring the 3D Standing Spine Posture from 2D Radiographs

This chapter has been published as peer-reviewed paper and is presented here with minor modifications.

Bayat, Amirhossein, Anjany Sekuboyina, Johannes C. Paetzold, Christian Payer, Darko Stern, Martin Urschler, Jan S. Kirschke, and Bjoern H. Menze. "Inferring the 3D standing spine posture from 2D radiographs." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 775-784. Springer, Cham, 2020. DOI: 10.1007/978-3-030-59725-2_75

Synopsis: This work presents a new deep learning based approach for synthesizing 3D shapes from 2D orthogonal images. An upright spinal pose (i.e., standing) under natural weight-bearing is crucial for such biomechanical analysis. 3D volumetric imaging modalities (e.g. CT and MRI) are performed in patients lying down. On the other hand, radiographs are captured in an upright pose but result in 2D projections. This work aims to integrate the two realms, i.e., it combines the upright spinal curvature from radiographs with the 3D vertebral shape from CT imaging for synthesizing an upright 3D model of the spine, loaded naturally. Specifically, we propose a novel neural network architecture working vertebra-wise, termed *TransVert*, which takes orthogonal 2D radiographs and infers the spine's 3D posture. We apply this model specifically to spinal radiographs. Applying this model to 2D radiographs enables us to generate 3D standing spinal postures applicable in biomechanical analysis of the spine. Training this model was a challenging task since the 3D ground-truths for radiographs are not available. Thus, we trained the model on synthetic data by simulating radiographs from CT images. However, the spinal CT scans available in the hospital are usually not the full field of view to decrease the scan size and save more disk space. Thus, the resulting 2D digitally reconstructed radiographs from CT images were not similar enough to real radiographs. We obtained large-scale datasets from lung scans where the lungs and ribs were available in the scan and created our synthetic dataset.

contributions of thesis author: algorithm design and implementation, generating the synthetic training data, experiment design and composition of manuscript.

4.1 Abstract

The treatment of degenerative spinal disorders requires an understanding of the individual spinal anatomy and curvature in 3D. An upright spinal pose (i.e. standing) under natural weight bearing is crucial for such bio-mechanical analysis. 3D volumetric imaging modalities (e.g. CT and MRI) are performed in patients lying down. On the other hand, radiographs are captured in an upright pose, but result in 2D projections. This work aims to integrate the two realms, i.e. it combines the upright spinal curvature from radiographs with the 3D vertebral shape from CT imaging for synthesizing an upright 3D model of spine, loaded naturally. Specifically, we propose a novel neural network architecture working vertebra-wise, termed *TransVert*, which takes orthogonal 2D radiographs and infers the spine's 3D posture. We validate our architecture on digitally reconstructed radiographs, achieving a 3D reconstruction Dice of 95.52%, indicating an almost perfect 2D-to-3D domain translation. Deploying our model on clinical radiographs, we successfully synthesise full-3D, upright, patient-specific spine models for the first time.

4.2 Introduction

A biomechanical study of spine and its load analysis in upright standing position is an active research topic, especially in cases of spine disorders [63]. Most common approaches for load estimation on the spine either use a general computational model of the spine for all patients or acquire subject-specific models from MRI or CT [64]. While these typical 3D image acquisition schemes capture rich 3D anatomical information, they require the patient to be in a *prone* or *supine* position (lying on one's chest or back), for imaging the spine. But, analysis of the spine's shape and vertebral arrangement needs to be done in a physiologically upright standing position under weight bearing, making 2D plain radiographs a *de facto* choice. A combination of both these worlds is of clinical interest to fully assess the bio-mechanical situation, i.e. to capture patient-specific complex pathological spinal arrangement in a standing position and with 3D information [65, 64, 66].

In literature, numerous registration-based methods have been proposed for relating 2D radiographs with 3D CT or MR images. In [66], the authors propose a rough manual registration of 3D data to 2D sagittal radiographs for the lumbar vertebrae. For the same purpose, in [67], manual annotations of the vertebral bodies are used as guideline for measuring the vertebral orientations in upright standing position. These methods are time and manual-labour-intensive and thus prone to error. Moreover, both these works use only the sagittal radiographs for vertebra positioning, while ignoring the coronal reformation which is a strong indicator of the spine's natural curvature. Aiming at this objective, [68] introduced an automatic 3D–2D spine registration algorithm, where the authors propose a multi-stage optimization-based registration method by introducing a metric for comparing a CT projection with a radiograph. However, this metric is hand-crafted, parameter-heavy, and is not learning-based, thus limiting its generalizability.

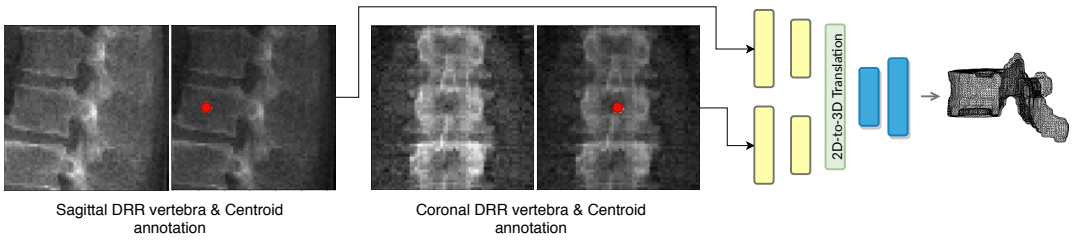


Figure 4.1: Overview of 2D image to 3D shape translation. The network inputs are 2D orthogonal view vertebrae patches and the centroid indicating the vertebra of interest.

In [69], the 3D shape of the spine is reconstructed using a biplanar X-ray device called 'EOS'. Hindering its applicability is the high device cost and the lack of its presence in a clinical routine. Recently the problem of reconstructing 3D shapes given 2D images have been explored using deep learning approaches. An approach closest to ours was proposed by Ying et al. [72], where they introduce a deep neural network to synthesize 3D CT images given orthogonal radiographs using adversarial networks. However, this model is highly memory intensive and fails to synthesise smaller anatomies like vertebrae in 3D. Moreover, it has been evaluated only on DRR, and its clinical applicability remains to be validated.

4.2.0.1 Motivation

The problem of 3D reconstruction of a spine in an anatomically upright position from 2D radiograph images relies on retrieving information from radiographs, which are 2D projections of a 3D object. Spine's sagittal reformation captures crucial information in the form of the vertebral body's and process' shape and its orientation around the sagittal (left-right) axis. However, its orientation around the cranio-caudal and anterior-posterior axes is obfuscated (cf. Fig. 4.1).

This information is available when combining sagittal with coronal reformations (or lateral with a.p. radiographs). Motivated by this, we propose a fully-supervised, computationally efficient, and registration-free approach combining sagittal and coronal 2D images to synthesise the vertebra's 3D shape model. Specifically:

- We introduce a novel FCN architecture for fusing orthogonal radiographs to generate 3D shapes.
- We identify an approach for training the network on synthetically generated radiographs from CT, being supervised by the CT's 3D vertebral masks.
- Validating our approach, we achieve dice score of **95.52%** on digitally reconstructed radiographs. We also successfully reconstruct 3D, patient-specific spine models on real clinical radiographs.

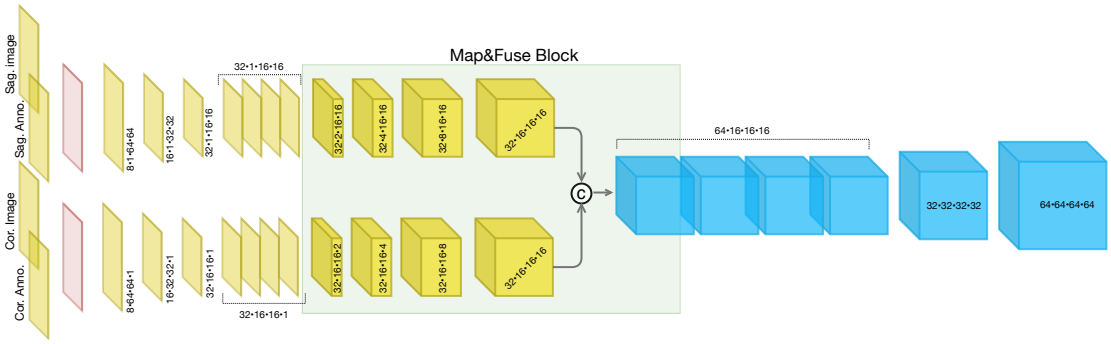


Figure 4.2: Architecture of *TransVert*. Our model is composed of sagittal and coronal 2D encoders (self-attention module in red), a 'map&fuse' block, and a 3D decoder.

4.3 Methods

Generating 3D shapes from 2D information is an ill-posed problem. For solving this, we utilize information from two orthogonal radiographs and an annotation on the vertebra of interest while relying on the shape prior learnt by the network.

4.3.1 TransVert: Translating 2D information to 3D shapes

The network performing 2D-to-3D synthesis needs to address the following requirements: First, it needs to appropriately combine information in the sagittal and coronal projections to recover 3D information. Second, recovering 3D shapes from 2D projections is inherently an ill-posed problem, requiring incorporation of prior knowledge. Lastly, the size of certain vertebra (towards the scan's periphery) is larger in radiographs compared to their true size due to the cone-beam of gamma-ray source. This effect should be negated when reconstructing the 3D model, i.e. the mapping should not be purely image-based. We address these requirements by proposing the *TransVert* architecture.

4.3.1.1 Overview

TransVert takes four 2D inputs, the sagittal and coronal vertebral image patches and their corresponding annotation images indicating the Vertebra of Interest (VOI). Denoting the 2D vertebral sagittal and coronal reformations by x_s and x_c, and their corresponding VOI annotation by y_s and y_c, we desire the vertebra's full-body 3D shape, $\mathbf{y}$, as a discrete voxel-map:

$$\mathbf{y} = G(x_s, x_c, y_s, y_c), \tag{4.1}$$

where G denotes the mapping performed by TransVert. In our case, the VOI-annotation image is obtained by placing a **disc of radius 1** around the vertebral centroid. In Section A.4, we analyze denser annotation choices (vertebral body and full vertebral masks). Ideally, training the TransVert mapping requires radiograph images and their

corresponding 'real world' 3D spine models. However, this correspondence does not exist and is, in fact, the problem we intend to solve. Thus, TransVert is trained on sagittal and coronal DRR constructed from CT images. It is supervised by the corresponding CT images' voxel-level, vertebral segmentation masks. As DRRs are similar in appearance to real radiographs, a DRR-trained TransVert architecture paired with a robust training regime, can be readily deployed on clinical radiographs.

4.3.1.2 Architecture

TransVert consists of three blocks: a 2D sagittal encoder, a 2D coronal encoder, and a 3D decoder. The three blocks are combined by a 'map&fuse' block. Refer to Fig. A.3 for a detailed illustration. The map&fuse block is responsible for *mapping* 2D representations of each the sagittal and coronal views into intermediate 3D latent representations followed by *fusing* them into a single 3D representation by channel-wise concatenation. This representation is then decoded into a viable 3D voxelized representation by the decoder. Note that the intermediate 3D representation is constructed from orthogonal views. Therefore, map&fuse block consists of anisotropic convolutions, with an anisotropy along the dimensions that need to be expanded. For example: the anterior-posterior dimension needs to be expanded for a coronal view. Consequently, the convolutional strides and padding directions are orthogonal for each of the view. At the network encoders' input, the vertebral images and VOI-annotations are combined using a self-attention layer. It was empirically observed that the attention mechanism yielded a better performance than a naive fusion by concatenating them as multiple channels.

4.3.1.3 Loss

Using solely a regression loss leads to converging to a local optimum where a mean (or median) shape is predicted, especially in the highly varying regions of the vertebra such as the vertebral processes. This is rectified by augmenting the loss with an adversarial component which checks the validity of a prediction at a global level. Therefore, TransVert is trained in a fully supervised manner by optimizing a combination of an ℓ_1 distance-based regression loss and an adversarial loss based on the least-squared GAN (LSGAN, [**?**]). Formally, the TransVert and the Discriminator combination is trained by minimizing the following losses:

$$\mathcal{L}_G = \alpha_G ||\mathbf{y} - G(x_s, x_c, y_s, y_c)||_1 + \alpha_D \left(D(G(x_s, x_c, y_s, y_c)) - 1\right)^2 \text{ and} \tag{4.2}$$

$$\mathcal{L}_D = (D(\mathbf{y}) - 1)^2 + D(G(x_s, x_c, y_s, y_c))^2, \tag{4.3}$$

where D represents the discriminator network and G represents the TranVert. α_G and α_D are weights of loss terms and fixed to $\alpha_G = 10$ and $\alpha_D = 0.1$. Note that $\mathbf{y}$ is binary valued containing $\{0, i\}$, where $i \in \{8, 9, \ldots 24\}$ denotes the vertebral index from T1 to L5. Forcing the network to predict the vertebral index implicitly incorporates an additional prior relating the shape to the vertebral index. Details about the discriminator

architecture and the adversarial training regime are provided in the supplemental material. The network is implemented with Pytorch framework on a Quadro P6000 GPU. It is trained till convergence using an Adam optimizer with initial learning rate is 0.0001.

4.4 Results

In this section, we describe the creation of DRRs, present an ablative study quantitatively analyzing the contribution of various architectural components, compare various VOI-annotation types, and finally deploy TransVert on real clinical radiographs.

4.4.1 Data

Recall that TransVert works with two data modalities: it is trained on DRRs extracted from CT images while being supervised by their corresponding 3D segmentation mask, and it is deployed on clinical radiographs.

4.4.1.1 CT data

We work with two datasets: a publicly available dataset for lung nodule detection with 800 chest CT scans [97], and an in-house dataset with 154 CT scans. In all, we work with $\sim 12k$ vertebrae split $5:1$ forming the training and validation set, reporting 3-fold cross validated results. Note that very few lumbar vertebrae are visible in [97] as it is lung-centred.
Data Preparation: The CT scans are segmented using [4] and the generated masks are validated by an experienced neuro-radiologist in order to consider only accurate ones for the study. These vertebral masks are used for supervision. Generation of the corresponding DRR is performed using a ray-casting approach [98], wherein a line is drawn from the radiation source (focal point) to every single pixel on the DRR image and the integral of the CT intensities over this line are calculated. Parameters for this generation include the radiation source-to-detector (= 180cm in this work) and the source-to-object distance (= 150cm here). Post the generation of the sagittal and coronal DRR, patches of size 64×64 are extracted around each vertebral centroid, constituting the image input to TransVert. The second input, viz. the VOI-annotation, can be extracted from the projected segmentation mask.

4.4.1.2 Clinical radiographs

We clinically validate TransVert on real long standing radiographs in corresponding lateral and anterior-posterior (a.p.) projections obtained in 30 patients. Acquisition parameters such as source-to-detector and source-to-object distances were similar to those used for DRR generation. Vertebral centroids needed for the VOI-annotations were automatically generated on both views using [99].

4.4.1.3 Data normalization

TransVert is trained on DRRs and tested on clinical radiographs. These data modalities have different intensity ranges, requiring normalization. We observe that z-score normalization works well, i.e. $\mathcal{I} = {}^{(\mathcal{I}-\mu_{\mathcal{I}})}/_{\sigma_{\mathcal{I}}}$, where $\mu_{\mathcal{I}}$ and $\sigma_{\mathcal{I}}$ are the mean and standard deviation of the image $\mathcal{I}$.

4.4.2 Experiments

We perform three sets of experiments validating our proposed approach, aimed at analysing the architectural aspects of TransVert, the data fed into it, and finally its applicability in a clinical setting. Note that a quantitative comparison with the ground truth can be performed only in experiments dealing with DRRs and CT images. Performance evaluation of various settings is compared by computing Dice coefficient and Hausdorff Distance between the predicted 3D vertebral mask and its ground truth from the CT mask.

4.4.2.1 Analysing TransVert's architecture

The proposed architecture for TransVert consists of the following architectural choices: fusion of sagittal and coronal views, anisotropic convolutions in the map&fuse block, a self-attention layer combining the image and the VOI-annotation, and finally, an adversarial component on the loss function. An ablative study over these components is reported in Table A.1. First, *do we need two views?*. For this, we evaluate the performance of a model that tries to reconstruct 3D shape from only a sagittal image. Next, *do we need anisotropic convolutions?*. For this, we compare two versions of map&fuse: one with a simple outer product for combining the orthogonal views (Naive View-Fusion) and one with the proposed anisotropic convolutions (TransVert). Observe that a simple fusion of views already outperforms a 'sagittal only' reconstruction. Also, anisotropic convolutions outperform fusion of views using outer-products. This can be attributed to the 2D-to-3D learning component involved in the latter. Lastly, *do we need the bells & whistles on top of TransVert?* Observe that incorporating the self-attention layer in the encoders and an adversarial training regime progressively improved performance, resulting in a Dice of 95.5% and a Hausdorff Distance of 5.11 mm. Fig. A.6 illustrates the 3D shape models reconstructed using the proposed architecture. Extracting a point cloud (with 2048 points) from these shapes, we also illustrate a point-wise Chamfer distance map. Observe that a vertebra's posterior region (vertebral process) is hardly visible in the image inputs. Despite this, TransVert is capable of recovering the process, albeit with a certain disagreement between the prediction and ground truth.

4.4.2.2 Analysing VOI-annotation type

Recall that alongside the image input, TransVert requires an auxiliary input indicating the vertebra of interest. We argue that a vertebral centroid suffices. In this study we show that our choice of vertebral centroid performs at a level comparable to a far

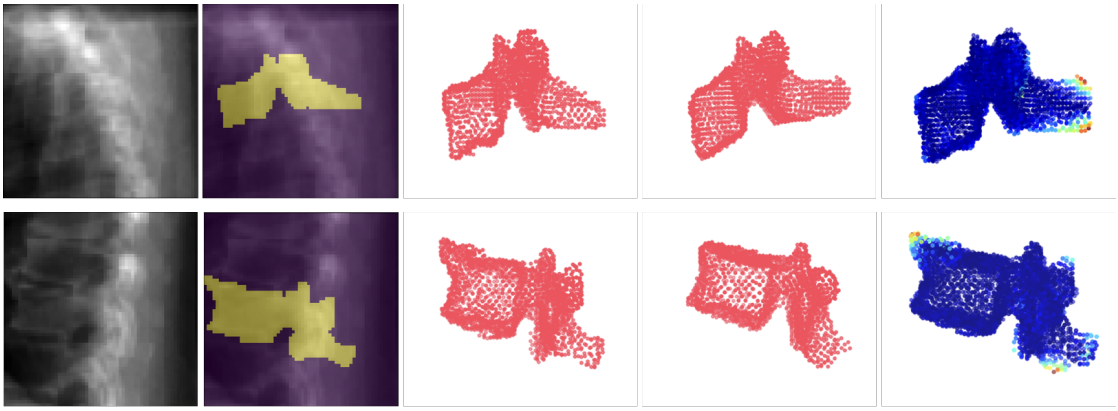


Figure 4.3: Shape modelling with TransVert on DRRs: First column indicates the image input. Second and third columns visualise the ground truth (GT) vertebral mask and the fourth visualises the predicted 3D shape model. Last column shows an overlayed Chamfer distance map between point clouds of GT and prediction.

Setup	Dice (%)	Hausdorff (mm)
Sagittal only	88.40	7.43
Naive View-Fusion (Outer Product)	92.59	6.45
TransVert	94.75	5.75
TransVert + Self Attn.	95.31	5.27
TransVert + SelfAttn + Adv.	95.52	5.11

Table 4.1: Architectural ablative study: The performance progressively improves with addition of each component. (Vertebral centroids are the VOI-annotations here.

denser full-vertebra annotation as reported in Table 4.2. We compare our centoids-to-vertebra (C2V) setup to two other, denser annotations: one where the vertebral body is annotated in the DRR (B2V) and one where the full vertebral body is annotated (Vertebra to Vertebra (V2V)).

As baseline, we include a setup without any VOI-annotation as an auxiliary input. Note that including the annotation input offers approximately 20% improvement in the mean Dice coefficient. Observe that a most dense V2V annotations and our C2V annotations perform comparably with only $< 1\%$ difference. Therefore, C2V is an obvious choice owing to the ease of marking centroids, more so because of existing automated labelling approaches.

Input	Dice (%)	Hausdorff (mm)
No annotation	76.44	14.74
V2V	96.24	4.18
B2V	95.67	4.95
C2V	95.31	5.27

Table 4.2: VOI annotation study: Performance drop from a denser (V2V) to a sparser annotation (C2V) is minor, while annotation effort decreases manifold.

4.4.2.3 2D-to-3D translation in clinical radiographs

TransVert works with individual vertebral images and their centroids. A 3D model of the spine can be constructed by stacking the predicted 3D vertebrae models at their corresponding 3D centroid locations. Vertebra's position along the axial and coronal axes is obtained from the sagittal reformation and its sagittal position from the coronal reformation. Fig. A.8 illustrates the results of this process. The top row visualises a 3D spine reconstruction based on 2D DRRs and compares it with the ground truth. More importantly, the bottom row depicts a successful deployment of TransVert in reconstructing the 3D, patient-specific posture of upright standing spine. Note that no 3D ground truth spine model exists for these cases. We visualise the 2D overlay of the segmentation on the radiographs, and the sagittal and coronal view of its 3D shape model, the former overlaid on the radiograph too. Observe that the 3D model's posture matches with that of the radiographs.

4.5 Conclusion

We propose TransVert, a novel architecture trained to infer a full-3D spine model from 2D sagittal and coronal radiographs and sparse centroid annotations. We identify an approach to train TransVert on DRRs in a fully-supervised manner. Along with an ablative study on TransVert's architectural components, we show a successful use case of deploying it on a real-world clinical radiograph.

4.5.0.1 Acknowledgements

This work was funded from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (GA637164–iBack–ERC-2014-STG).

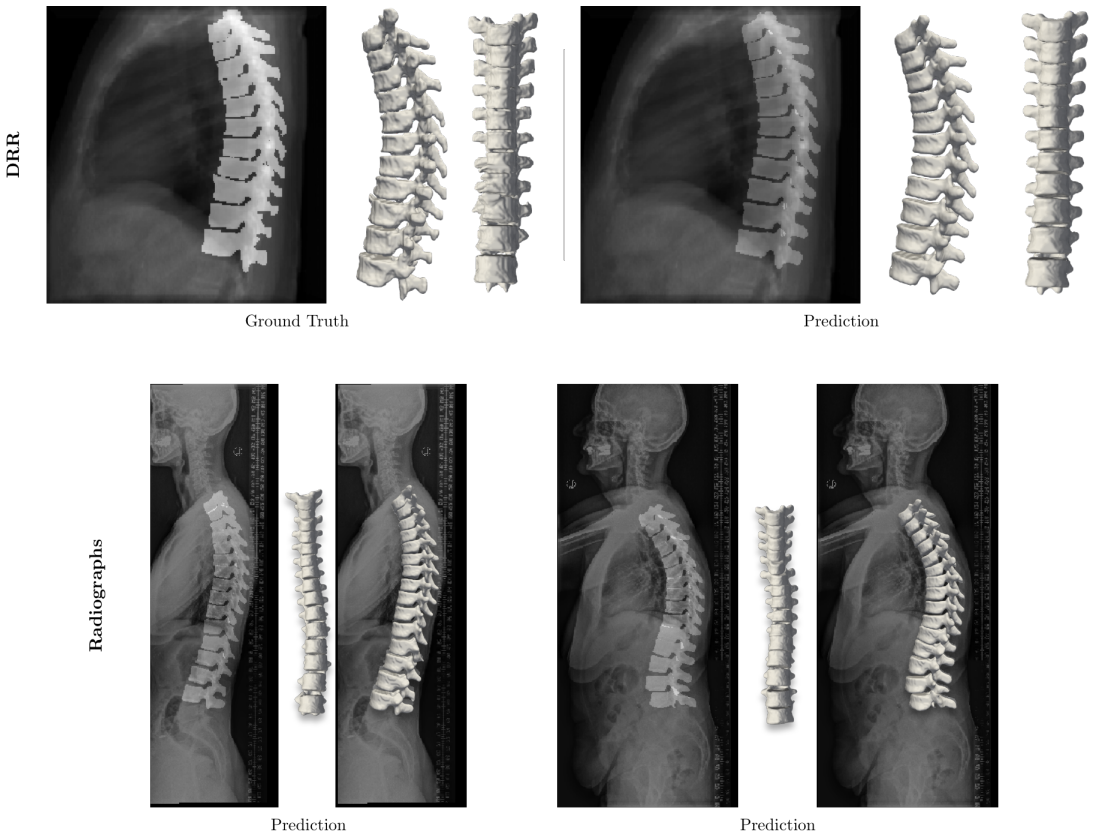


Figure 4.4: Full 3D spine models: (Top row) Comparison of a DRR-based spine model reconstruction with its CT ground truth mask. (Bottom row) 3D patient-specific spine models constructed from real clinical radiographs.

5 Cranial Implant Prediction using Low-Resolution 3D Shape Completion and High-Resolution 2D Refinement

This chapter has been published as peer-reviewed paper and is presented here with minor modifications.

Bayat, Amirhossein, Suprosanna Shit, Adrian Kilian, Jürgen T. Liechtenstein, Jan S. Kirschke, and Bjoern H. Menze. "Cranial Implant Prediction Using Low-Resolution 3D Shape Completion and High-Resolution 2D Refinement." In Cranial Implant Design Challenge, pp. 77-84. Springer, Cham, 2020. DOI: 10.1007/978-3-030-64327-0_9

Synopsis: This work introduces a novel approach for designing patient-specific cranial implants using deep neural networks. Due to GPU memory limitations, we devised a solution composed of two sub-networks functioning on 3D and 2D images with low and high resolutions. Our model performs on 3D data in low resolution since a 2D model lacks a holistic 3D view of both the defective and healthy skulls. The first sub-network is designed to complete the shape of the downsampled defective skull. The second sub-network upsamples the reconstructed shape slice-wise. We train both the 3D and 2D networks in tandem in an end-to-end fashion, with a custom hierarchical loss function. Our proposed solution accurately predicts a high-resolution 3D implant in the challenge test case in terms of dice-score and the Hausdorff distance.

contributions of thesis author: algorithm design and implementation, experiment design and composition of manuscript.

5.1 Abstract

Designing of a cranial implant needs a 3D understanding of the complete skull shape. Thus, taking a 2D approach is sub-optimal, since a 2D model lacks a holistic 3D view of both the defective and healthy skulls. Further, loading the whole 3D skull shapes at its original image resolution is not feasible in commonly available GPUs. To mitigate these issues, we propose a fully convolutional network composed of two subnetworks. The first subnetwork is designed to complete the shape of the downsampled defective skull. The second subnetwork upsamples the reconstructed shape slice-wise. We train both the 3D and 2D networks in tandem in an end-to-end fashion, with a hierarchical loss function. Our proposed solution accurately predicts a high-resolution 3D implant in the challenge test case in terms of dice-score and the Hausdorff distance.

5.2 Introduction

Cranial implant design is a crucial task for clinical planning of cranioplasty [100]. Previous works mainly rely on freely available CAD tools for cranial implant design [101, 102, 103, 104]. The time requirements and need for expert intervention for these approaches are a major hindrance for fast and in-prompt deployment. The AutoImplant challenge aims to look for simple and easy-to-use automatic solution that can accurately predict cranial implants. Keeping this in mind, we tailor our proposed solution to best fit the requirements of clinicians.

Previous literature [105] tend to exploit the geometric symmetry and predict cranial implant based on the unaffected skull region. Nevertheless, this results in a suboptimal solution, since the human skull is not perfectly symmetric in reality. These solutions also fall short when the implant is not exclusively in one hemisphere. Morais et al. [106] used a deep 3D encoder-decoder [2, 87, 107] network to reconstruct the incomplete skull in low-resolution space. While the low-resolution space facilitates faster processing, the quality of the reconstruction lacks minute local anatomical detail. In the Autoimplant baseline paper [108], a similar approach is taken where the authors first localize the defective region in the skull and then predict the implant using an encoder-decoder network. While this pipeline is suitable for modular design of accurate defective region detection and implant prediction, the network is not end-to-end trainable, and thus any error during the first stages would penalize the implant prediction.

In this approach, we try to alleviate this by relying on coarse-scale implant prediction in 3D followed by fine-scale enhancement of the predicted implant. We identify that 3D is most suitable to predict the implant since anatomical consistency is best captured by a 3D receptive field compared to any local 2D slices. However, to reduce the memory and computational power, we first predict the implant in a down-sampled defective skull. Subsequently, we enhance the predicted implant slice wise by a 2D decoder network. Thus our solution becomes end-to-end trainable and also is efficient at the same time for the high-resolution implant prediction task.

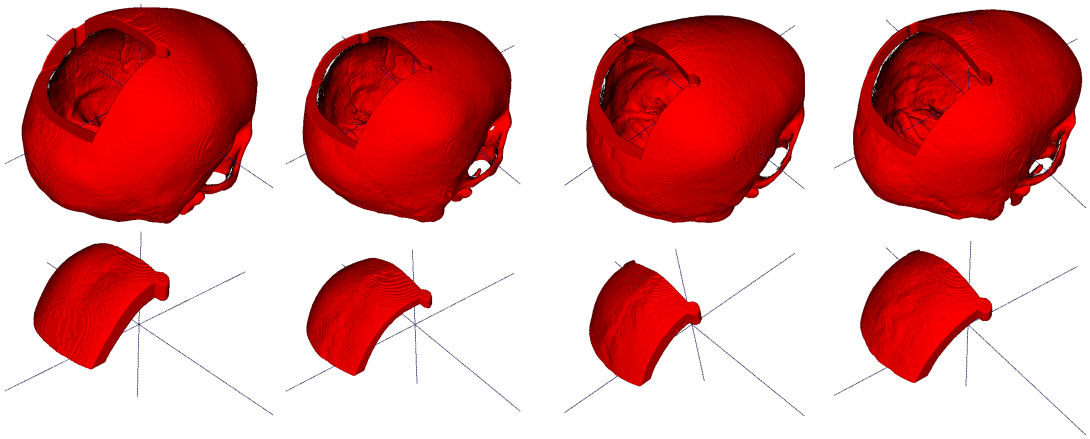


Figure 5.1: Few Training sample: The first row depicts rendered 3D volumes of four randomly selected defective scan from the training dataset. The second row shows the corresponding ground truth cranial implant.

5.3 Method

The dataset is created by artificially generating the defect in the scan [108]. Thus the original skull would be the ground truth for the implant prediction task. We leverage this availability of the target label and cast the implant prediction task as a supervised volumetric reconstruction task. At the core of our method lies a 3D encoder-decoder network. This network takes the low-resolution defective skull as input and predicts a low-resolution implant at the output. We argue that the implant prediction task lies in a lower-dimensional manifold since the key properties to predict implant are the inner and outer surface consistency. Hence, a down-sampled input space is sufficient for a coarse-scale identification of the implant region. A simple element-wise subtraction of the reconstructed skull and the input will produce the desired implant. This approach is in line with the shape completion literature [109, 110, 111, 48, 112]. Next, we need to upsample the predicted implant, which can be done in several ways. Classical approaches, such as spline-based interpolation, can be a simple choice. Alternatively, a decoder network proved to be superior in the super-resolution task [113, 114]. Hence, we incorporate a second module in our method, a 2D up-sampler. This up-sampler takes selected axial slices during training and predicts the up-sampled version of it. To be able to train the both the network jointly and also fit the data in the GPU memory, we select N random slices out of the reconstructed shape and select the corresponding slices from the original scale Ground Truth. The error between the predicted slice and the ground truth skull is used to train the 2D decoder. The high-resolution reconstruction error, along with the 3D shape completion error, contributes to the training of the 3D encoder-decoder.

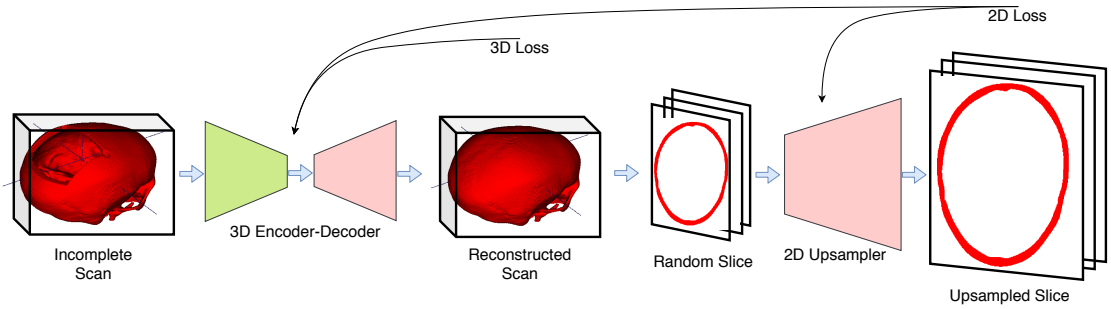


Figure 5.2: Schematic overview of our proposed pipeline for predicting the cranial implant. The downsampled defective scan goes through an encoder-decoder based shape completion network. During training, N number of random reconstructed skull goes through a second decoder network for high-resolution reconstruction. For the 3D shape completion, we use a volumetric ℓ_1 norm, and for the 2D refinement task, we use summation of 2D ℓ_1 loss.

5.3.1 Network Architecture & Loss Function

In the following, we describe the architecture of two subnetworks in our model and the loss functions used to train the model.

5.3.1.1 3D Encoder-Decoder:

Encoder-decoder type network has been previously used in bio-physical simulation [115], image segmentation [116] etc. Our 3D network has three sequential components, such as an encoder, bottleneck, and a decoder. The encoder further compresses the input signals into a more compact representation, which is processed in the bottleneck unit to extract useful features. These features go through the decoder part to reconstruct the complete skull. The complete architecture is as follows:
$IN_1 \rightarrow CN^1_{64} \rightarrow CN^2_{64} \rightarrow CN^2_{64} \rightarrow RB_{64} \rightarrow RB_{64} \rightarrow RB_{64} \rightarrow RB_{64} \rightarrow TC^2_{64} \rightarrow TC^2_{64} \rightarrow C^1_1 \rightarrow OUT_1$
where IN_1 and OUT_1 is input and output volume respectively with single channel, $CN^s_{\#ch}$ is convolution with stride s and output channel $\#ch$ followed by batch norm and ReLU, $TC^s_{\#ch}$ is transposed convolution with stride s and output channel $\#ch$ followed by batch norm and ReLU, $RB_{\#ch}$ is residual block consists of two successive unit of convolution with stride 1 and output channel $\#ch$ followed by instance norm and ReLU, and $C^s_{\#ch}$ is convolution with stride s and output channel $\#ch$ followed by sigmoid. Note that all convolution and norm layers described here are 3D.

5.3.1.2 2D Decoder Upsampler:

The 2D upsampler network consists of four residual blocks, followed by the nearest neighborhood upsampling layer and a final convolution layer. The residual blocks refine the low-resolution reconstructed scans to incorporate anatomical consistency, which aids precise high-resolution skull at the output. We concatenate the corresponding slice of the defective scan along with the reconstructed scan and pass it as an input to the

2D upsampler. This helps to correct any location-wise mismatch in the 3D shape-completion task. Borrowing a few notations defined in the previous paragraph, the complete architecture is given below:
$IN_2 \rightarrow CN_{64}^1 \rightarrow SE_{64} \rightarrow RB_{64} \rightarrow SE_{64} \rightarrow RB_{64} \rightarrow SE_{64} \rightarrow RB_{64} \rightarrow SE_{64} \rightarrow RB_{64} \rightarrow NN_{64}^{sqrt(512/180)} \rightarrow NN_{64}^{sqrt(512/180)} \rightarrow C_1^1 \rightarrow OUT_1$
where $SE\#ch$ is 'squeeze and excitation' layer and $NN_{\#ch}^s$ is Nearest Neighborhood (NN) upsample with scale factor s and output channel $\#ch$ followed by instance norm and ReLU. Note that all convolution and norm layers described here are 2D.

5.3.1.3 Loss Function:

Let's denote the ground truth data at original scale as I_G, downsampled ground truth data I_g, defective 3D volume at original scale as I_D, downsampled defective 3D volume as I_d, the functional form of the 3D encoder-decoder network as $S()$, and the functional form of the 2D upsampler network as $U()$ respectively. The cranial implant is predicted as follows:

$$\text{Cranial Implant } = U(S(I_d)) \setminus I_D \tag{5.1}$$

where $\setminus$ denotes set difference. The total loss function of our method is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{3D} + \mathcal{L}_{2D} \tag{5.2}$$

$$\mathcal{L}_{3D} = \|S(I_d) - I_g\|_{\ell_1} \tag{5.3}$$

$$\mathcal{L}_{2D} = \sum_{i \in \Omega} \|U(S(I_d)^i) - I_G^i\|_{\ell_1} \tag{5.4}$$

where Ω is the set of random slices

5.3.2 Implementation

We realize our model in PyTorch. We trained the networks with Adam optimizer and a learning rate of 0.0001. We used an Nvidia Quadro P6000 GPU. The batch size for the 3D network was 1, so one volume per iteration. **We downsampled the original 3D volume by a factor of $\frac{512}{180}$ in all dimension because that is the largest 3D volume we can fit in our GPU along with the 2D decoder module. The downsampled 3D volume is zero-padded in the z-dimension to make it** $180 \times 180 \times 180$**.** After predicting the completed 3D shape in low resolution, we sample 10 slices randomly along the Z-axis and concatenate them channel-wise with the downsampled corresponding slice from the defective skull and feed them to the upsampler decoder. We can't fit the entire volume with the original scale in the memory, so we have to select 2D slices. In order to avoid overfitting, we select the slices randomly. It is important to note that, after downsampling the volume with a $\frac{512}{180}$ scaling factor, every 3 slices along Z-axis in the original scale correspond to 1 slice in the downsampled volume. Thus, after reconstructing the 3D shape, we have a set of selected slices using random indices and three sets of [random indices/0.35], [random indices/0.35]+1 and [random indices/0.35]+2. We select the corresponding slices from the defective scan and downsample them in 2D to

be concatenated with the slices from the predicted shape. Thus, the batch size for the upsampler decoder network is 30.

5.3.2.1 Inference:

For inference, similarly, a downsampled volume is fed to the network, and it is reconstructed in low resolution using the first sub-network. After that, all of the slices along the Z-axis are fed to the upsampler decoder, one-by-one, and stacked in volume to reconstruct the shape in 3D. Subsequently, we subtract the defective input scan from the high-resolution reconstructed scans to estimate the cranial implant. Finally, as a post-processing step, we erode and dilate the segmentation consequently with a sphere structure with a radius of 2 to remove the noise. Subsequently, we select the largest component in the segmentation map, using connected component analysis.

5.4 Experimental Results

Table 5.1: Our score on the validation dataset

Method	**Dice**	**HD-distance**
Ours (Transposed Conv)	0.8363	10.6570
Ours (NN Upsampling)	**0.9358**	**7.6100**

We work with 100 data samples split 5 : 1 forming the training and validation set. We validate our approach by comparing the constructed implants to the ground truth, using the Dice score and the Hausdorff distance. The validation results are presented in Table 5.1. **We experimented with two variations of our 2D decoder model. In the first case, we trained the original decoder, and in the second case, we replace the NN upsampling layers of the 2D decoder with transposed convolution layers.** We observe that the 2D decoder with NN upsampling layers performs significantly better than the decoder with transposed-convolution layers. We attribute this to the over parameterization during the upsampling step. Since the image is binary in nature, the nearest neighborhood upsampling layer is sufficient for this task.

Table 5.2: Our score on the 100 test cases.

Method	**Dice**	**HD-distance**
Baseline [108]	0.8555	5.1825
Ours (NN Upsampling)	**0.8957**	**4.6019**

Finally, we tested our model on the challenge test set and report the results for cases $000 \sim 099$ in Table 5.2. Our model outperforms the baseline method proposed in [108] both in the Dice score and Hausdorff distance. We did not report the results on cases

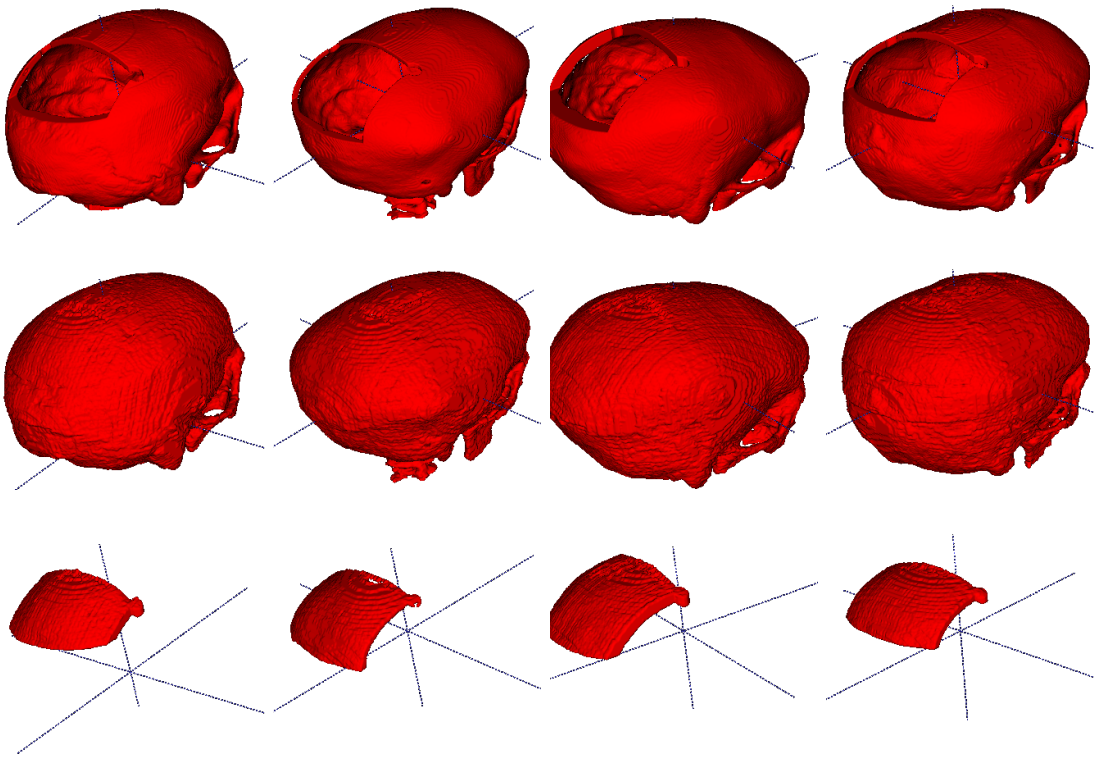


Figure 5.3: Qualitative results: The first row depicts four rendered 3D volume of defective scan from the test dataset. The second row shows the reconstructed skull by our method of the corresponding defective skull. The third row is the corresponding cranial implant predicted by our method. We observe that our method generalizes well and accurately reconstruct the skull to predict the cranial implants.

$100 \sim 109$, since the location of the defect is very different from the training set, and the model could not predict the implant. Fig. 5.3 shows the qualitative results of randomly selected scans from the test data set. Visual inspection also confirms that our model estimates accurate cranial implants for these cases. The source code of our model is accessible from https://github.com/mlentwicklung/autoimplant.

5.5 Conclusion

We provide an efficient and compact solution for the AutoImplant 2020 challenge, which is suitable for fast and easy deployment. Our key innovation is the incorporation of a two-stage reconstruction policy, where the first stage predicts a coarse-scale implant, and the second stage super-resolve it to a high-resolution one. We achieve accurate implant prediction on the validation dataset. Our model is end-to-end in the high-resolution space and thus can serve as a baseline for developing more complex models aiming to better learn the anatomically invariant implant prediction.

Acknowledgement

Amirhossein Bayat is supported by the European Research Council (ERC) under the European Union's 'Horizon 2020' research & innovation programme (GA637164–iBack–ERC–2014–STG). Suprosanna Shit is supported by the Translational Brain Imaging Training Network (TRABIT) under the European Union's 'Horizon 2020' research & innovation program (Grant agreement ID: 765148).

6 Conclusion & Outlook

This publication-based dissertation presents methods for processing the 2D and 3D spinal scans; it also covers some techniques for predicting cranial implant design. In total, this work includes three publications and one manuscript currently being under review for publication. Chapters 3 to 5 are self-contained and in their original form. Appendix A provides additional unpublished work which might evolve in the future. This chapter provides an overview of the previous chapters and concludes with the directions for future work.

This thesis's contributions are related to deep learning-based shape completion, vertebral detection and labeling in X-ray images, 2D/3D shape inference, predicting 3D cranial implants, 2D/3D registration, prior based 3D shape reconstruction and registration.

We presented a deep neural network for detecting and labeling the vertebra on radiographs in Chapter 3. Detecting and labeling vertebra in the radiograph is a challenging task due to the heavy tissue overlay and noise. Most of the works for vertebra labeling were on CT images, and our work was one of the first methods proposed for radiographs. We proposed a new model architecture, robust to vertebra occlusion, by enforcing the spinal shape.

In Chapter 4 presents our work on deep learning based inference of 3D spinal postures from 2D radiographs. An upright spinal pose (i.e., standing) under natural weight-bearing is crucial for such biomechanical analysis. 3D volumetric imaging modalities (e.g. CT and MRI) are performed in patients lying down. On the other hand, radiographs are captured in an upright pose but result in 2D projections. This work aims to integrate the two realms, i.e., it combines the upright spinal curvature from radiographs with the 3D vertebral shape from CT imaging synthesizing an upright 3D model of the spine, loaded naturally. Specifically, we propose a novel neural network architecture working vertebra-wise, termed *TransVert*, which takes orthogonal 2D radiographs and infers the spine's 3D posture.

In Chapter 5 we presented a deep neural architecture for designing a cranial implant. Due to GPU memory limitations, we devised a solution composed of two subnetwork functioning on 3D and 2D images with low and high resolutions. Our model performs on 3D data in low resolution since a 2D model lacks a holistic 3D view of both the defective and healthy skulls. The first subnetwork is designed to complete the shape of the downsampled defective skull. The second subnetwork upsamples the reconstructed shape slice-wise. We train both the 3D and 2D networks in tandem in an end-to-end fashion, with a custom hierarchical loss function. Our proposed solution accurately predicts a high-resolution 3D implant in the challenge test case in terms of dice-score and the Hausdorff distance.

In continuation of our work form Chapter 4 we explored the idea of incorporating shape priors into the deep neural architecture to generate the 3D standing spine posture from 2D orthogonal spinal radiographs. To increase the robustness to noise and tissue overlay in radiographs, we introduced an anatomy-aware model for reconstructing the spinal postures in 3D. Our proposed model is informed of the spinal curve and each vertebral type shape by integrating shape priors into the network architecture. Our model learns to deform the provided shape templates to generate 3D shapes. Thus, during the training phase, the network estimates the deformation field required to register the shape templates to the target shapes. The model comprises different subnetworks for encoding the input data and decoder subnetworks to estimate the deformable and affine deformation fields for individual vertebra. The results of this work will be published as a journal paper in the future.

As in many other fields, recent deep learning progress has strongly influenced the latest medical imaging approaches. This influence includes registration tasks as well. Traditionally, the registration process is defined as an optimization problem, and the deformation parameters are estimated in an iterative procedure, which makes it time-intensive and prone to error. In our approach, once the model is trained, the registration is conducted in a single forward pass with higher accuracy and robustness.

Deep convolutional networks have demonstrated remarkable performance on medical imaging, while for shape processing tasks, working in voxel space and applying convolutional networks might be sub-optimal.

The emerging field of research in geometric deep learning aims at transferring the concepts of deep learning to non-euclidean domains. Non-euclidian domains include learning on graphs, meshes[42, 117] and variants of which are starting to find application in medical image computing such as automatic implant design and organ shape processing.

Due to the ever-increasing number of parameters in deep neural networks, the demand for training data to feed the networks is swelling. However, providing annotated data, especially for segmentation tasks, is not feasible. Devising weakly supervised learning and self-supervised learning to take advantage of unlabeled or sparsely labeled data is important.

A Appendix: Anatomy-aware Inference of the 3D Standing Spine Posture from 2D Radiographs

Unpublished work presented as part of the thesis.

A.1 Abstract

The balance of the spine is an important factor for the development of spinal degeneration, pain and the outcome of spinal surgery. It must be analyzed in an upright, standing position to ensure physiological loading conditions and visualize load-dependent deformations. Despite the complex 3D shape of the spine, this analysis is currently performed using 2D radiographs, as all frequently used 3D imaging techniques require the patient to be scanned in a prone position. To overcome this limitation, we propose a deep neural network to reconstruct the 3D spinal pose in upright standing position, loaded naturally. Specifically, we propose a novel neural network architecture, that takes orthogonal 2D radiographs and infers the spine's 3D posture using vertebral shape priors. In this work, we define vertebral shape priors using an atlas and a spine shape prior, incorporating both into our proposed network architecture. We validate our architecture on digitally reconstructed radiographs, achieving a 3D reconstruction Dice of 0.95, indicating an almost perfect 2D-to-3D domain translation. Validating the reconstruction accuracy of a 3D standing spine on real data is infeasible due to the lack of a valid ground truth. Hence, we design a novel experiment for this purpose, using an orientation invariant distance metric, to evaluate our model's ability to synthesize full-3D, upright, and patient-specific spine models. We compare the synthesized spine shapes from clinical upright standing radiographs to the same patient's 3D spinal posture in the prone position from CT.

A.2 Introduction

A biomechanical load analysis of the spine in an upright standing position is highly warranted in various spine disorders to understand their cause and guide therapy [63]. Typical approaches for load estimation either use a computational shape model of the spine for all patients or obtain a subject-specific spine model from a 3D imaging modality such as magnetic resonance imaging (MRI) or computed tomography (CT) [64]. However, eventhough MRI and CT images can capture 3D anatomical information, they need

the patient to be in a *prone* or *supine* position (lying flat on a table) during imaging. However, to analyze the spinal alignment in a physiologically upright standing position under weight bearing, orthogonal 2D plain radiographs (as depicted in Fig. A.1) are the *de facto* choice. A combination of both these worlds is of clinical interest to fully assess the true bio-mechanical situation, i.e. to capture the patient-specific complex pathological spinal arrangement in a standing position with full 3D information [65, 64, 66].

As much spatial information is lost when projecting a 3D object in only two 2D planes, a random object cannot reliably be reconstructed from two orthogonal projections. However, the spine follows strong anatomical rules, that are repeated only with slight variations in any patient. Typical projections, i.e. lateral and a.p. radiographs, cover most of these variations, both on a local (per vertebra) and global (overall spinal alignment) level. Motivated by this, we propose a fully-supervised, computationally efficient, and robust approach combining sagittal and coronal 2D images to synthesise each vertebra's 3D shape model by deforming shape templates, forming a complete patient specific spine model.

Training a model for 3D shape synthesis from 2D images, requires 3D supervision. However, for 2D clinical radiographs we do not have the corresponding 3D shape of spine. Thus, we introduce a training approach using synthetically generated radiographs from 3D CT, with full supervision from the CT's 3D vertebral masks.

Furthermore, to stabilize the model, we incorporate spinal and vertebral shape priors, and train our model to estimate the deformation field to warp an atlas and generate the 3D shape [118]. In machine learning, shape priors help to reduce the search space of possible solutions, improving the accuracy and plausibility of solutions [23]. Priors are particularly effective when data are unclear, corrupt, with low signal-to-noise ratio or when training data are scarce [23]. As we apply our model on clinical radiographs subject to noise and heavy tissue overlay, incorporating shape priors into the model can enhance robustness against such artefacts in real-world data. As the task is to synthesize not only individual vertebral shapes, but also the complete spinal alignment, we show that additionally including a spine shape prior as global information improves the results.

As the position of the patient differs in prone 3D imaging and upright radiographs, the orientation of the generated radiograph-based 3D vertebrae are different from the CT scan of the same patient. Thus, to quantitatively compare the vertebral shapes of the model and ground truth, we need a metric invariant to the orientation for the comparison of the generated vertebral shapes to those from CT scan of the same patient. For this purpose we use an orientation-invariant metric.

Specifically:

- We introduce an anatomy-aware deep neural network for fusing orthogonal radiographs to generate a 3D spine model. We define this 3D shape synthesis as a hybrid registration problem in which the network estimates vector fields to deform vertebral shape templates to achieve shape synthesis.

- Validating our approach using 3 different metrics, we achieved scores of **0.95**, **5.70 mm** and **0.17** for Dice, Hausdorff distance and normalized weighted spec-

tral distance (nWESD) on digitally reconstructed radiographs. We successfully reconstructed 3D, patient-specific spine models on real clinical radiographs.

- Validating the success in a clinical setting required a novel experimental setup. We designed an experiment to compare the results on clinical radiographs to the vertebral shapes from the same patient's 3D scan. Since the two spine postures are dissimilar, the orientation of corresponding individual vertebrae are different in each posture. Thus, we use nWESD as a metric which is invariant to rigid transformations, to conduct a vertebra-by-vertebra comparison, and achieve the nWESD score of 0.13.

A.2.1 Related Work

The literature offers a wealth of pre-existing registration-based methods [119, 120, 77, 27, 121, 122] for relating 2D radiographs to 3D CT or MR images. In [66], the authors employ coarse manual registration for aligning 3D data to 2D sagittal radiographs for the lumbar vertebrae. Similarly, in [67], manually annotated vertebral bodies on 2D radiographs are used to measure the vertebral orientations in an upright (standing) position. Such solutions are time-consuming, labor-intensive, and are vulnerable to error. Note that both the works only employ sagittal reformations to position the vertebrae. They ignore the coronal reformation contains significant information about the spine's natural curvature, especially in abnormal cases. Other approahes for this purpose introduce an automated 3D–2D spine registration algorithm [68], [123], wherein authors suggest a multi-stage registration method by introducing a comparison metric for a CT projection and a radiograph. Since this metric is parameter-heavy and hand-crafted, the generalizability and inference speeds of these methods are limited. In [69], the 3D shape of the spine is reconstructed using a biplanar X-ray device called 'EOS'. The advantage of the system is the low radiation dose required and that both projections are acquired simultaneously, allowing for a direct spatial correspondence between the two planes. Hindering its applicability is the high device cost and thus the lack of its presence in clinical routine.

The problem of reconstructing 3D shapes from 2D images has recently been explored using deep learning methods. For example, in [72], adversarial training is used to synthesize 3D CT images given to orthogonal radiographs. For a scale of spinal scans, this method is memory intensive and also fails to synthesize smaller 3D anatomies such as the vertebrae. Note that this method has been evaluated on digitally reconstructed radiographs (DRR) only, thus requiring further validation for its clinical usage. In [73] and [74], the authors design a model to generate a 3D shape given multiple arbitrary view 2D images. However, the input images include only the object of interest without background, which is not applicable to medical images like spinal radiographs. In our previous work on inferring the 3D standing spine posture [124], we introduce a new deep neural network architecture termed TransVert, to combine the 2D orthogonal image information and reconstruct a 3D shape. Since the vertebrae are heavily occluded by soft tissue and ribs in the radiographs, the reconstruction is challenging and in some cases the

output shape is far from a vertebral shape. The idea of incorporating shape priors into the neural network models have been explored by [23, 75, 76, 77, 125]. However, for our application, considering only the vertebral shape priors when reconstructing the spine posture is not enough. Most of the ideas proposed for using shape priors in deep neural networks are for a single object, while the spine shape is more complicated as it is made of multiple objects (vertebrae) connected to each other and deforming subject to some constraints from human anatomy. Thus, defining a prior for explaining intervertebral constraints and the spinal shape is crucial here.

A few works have modeled the spinal shape [126]. For instance, in [78] the authors proposed an automatic framework that segments vertebrae from arbitrary CT images with a complete spine model. They first scanned a commercially available plastic phantom to generate the the template. Next, they manually registered it to ten actual full spine scans. Authors learned a statistical form model of the spine in [79] by independently studying three models for cervical, thoracic and lumbar regions. Thus, their models do not learn the shape correlations across the full spine. [126]

Other probabilistic models, such as probabilistic atlas [80], graph models [81], Hidden Markov Models [82] and hierarchical models [83, 84, 85] have also been proposed. For instance in [80] authors 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. In [81] the authors proposed a probabilistic graphical model for the location and identification of the vertebrae in MR images. In both cases full spines were observed at training time. But none of the works mentioned above are incorporated in a machine learning model for reconstructing 3D shape of spine from 2D images.

A.3 TransVert+: From 2D images to 3D shapes

Our network estimates the vector field that deforms a discrete atlas to the desired vertebral shape. As a final goal we require the network to generate a 3D posture of the spine given two orthogonal 2D radiographs. For synthesizing 3D data from 2D information, the the following requirements are desired: First, for efficient recovery of 3D shape information from sagittal and coronal projections, the network needs to integrate the information from these projectins appropriately. Second, recovering 3D shapes from 2D projections is inherently an ill-posed problem, requiring incorporation of prior knowledge. This knowledge includes vertebral shapes and the shape of the spine (spinal curvature). We incorporate a spine atlas as a shape prior into our network architecture, to enforce vertebral shape constraints. We also define the spinal curve by atlas vertebral centroids, for enforcing the shape of the spine. In addition to image data, we include the vertebral labels as an additional annotation attached to the network input. Fig. A.2 depicts orthogonal input image patches with corresponding centroid annotations, indicating the vertebra of interest in that patch. We take a registration approach for shape synthesis and propose the *TransVert+* network architecture.

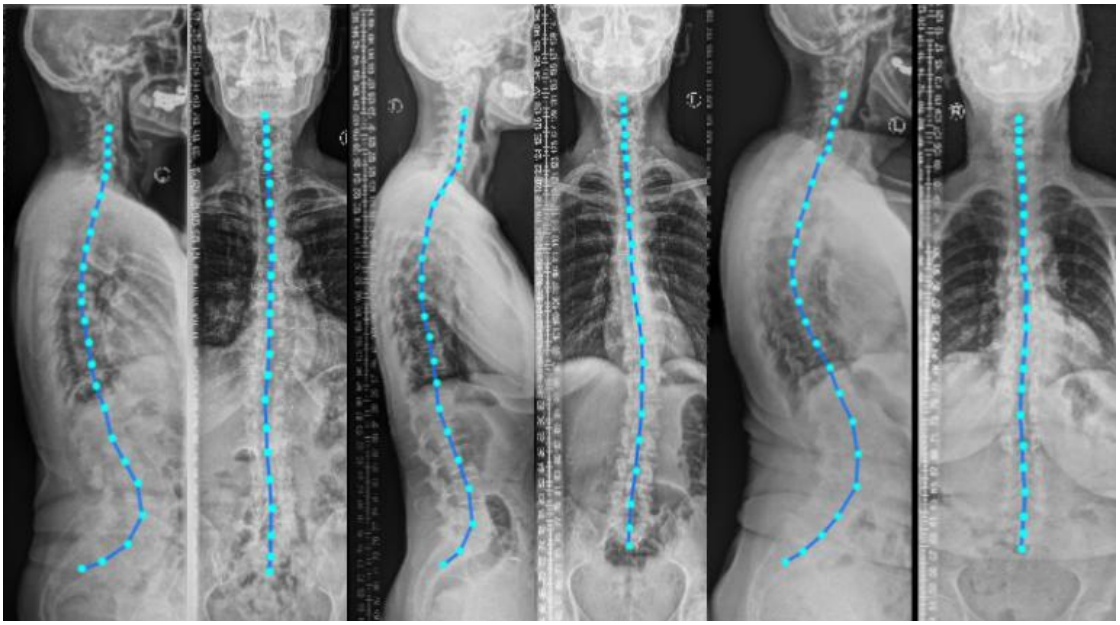


Figure A.1: Lateral and anterior-posterior (a.p.) view radiographs of 3 patients with spinal curvature annotation. Considering the vertebral centroid coordinates and spinal curvature facilitates determining the scale and the vertebral orientation.

A.3.1 Overview

TransVert+ takes five inputs, of which four are 2D images of the sagittal and coronal vertebral image patches (x_s and x_c) and their corresponding annotation images (y_s and y_c) indicating the vertebra-of-interest (VOI). In our case, the VOI-annotation image is obtained by placing a **disc of radius 1** around the vertebral centroid. The image patches and the corresponding annotation images are illustrated in Fig. A.2. The fifth input is a vector of floats denoting the coordinates of the vertebral centroids ($\vec{C}^{\,v}$) in a global spine coordinate system, for providing the model with a holistic view of the spinal curve for more consistency. We desire a function G that outputs the vertebra's full-body 3D shape, $\mathbf{y}$, which is represented as a discrete voxel-map by deforming the vertebral shape templates (y_t):

$$\mathbf{y} = G(x_s, x_c, y_s, y_c, \vec{C}^{\,v}) \circ y_t. \tag{A.1}$$

We formulate this shape synthesis as a registration problem in which vertebral shape templates are deformed to obtain desired shapes. This includes both global affine transformations (scaling and 3D rotation for each vertebra, considering the global spinal shape) and local deformations on the vertebral surface. We denote them as two subtasks,

$$G(x_s, x_c, y_s, y_c, \vec{C}^{\,v}) = G_a + G_d, \tag{A.2}$$

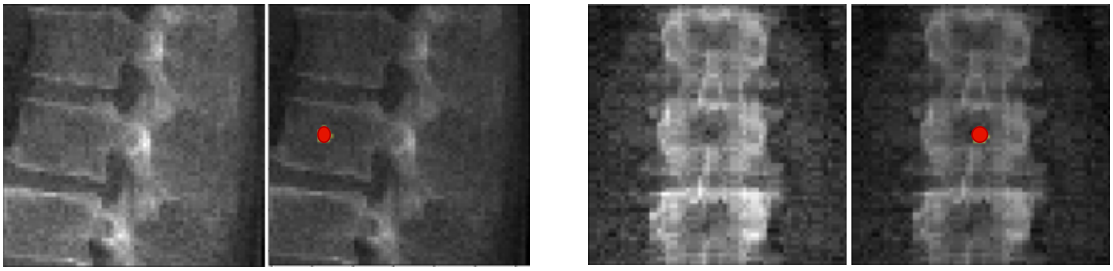


Figure A.2: Vertebral image patches with corresponding annotations. The network inputs are 2D orthogonal view vertebrae patches and the centroid indicates the vertebra of interest.

$$G_a = AffineDecoder(x_s, x_c, y_s, y_c, \vec{C}^v), \tag{A.3}$$

$$G_d = DeformableDecoder(x_s, x_c, y_s, y_c), \tag{A.4}$$

where G denotes the mapping performed by TransVert+. The transformation G is separated into affine (G_a) and deformable (G_d) transformations. When inferring the affine transformation the entire spinal shape is considered, while for inferring the deformable transformation individual vertebral shape features are taken into account. We will elaborate more on this in the next sections.

Ideally, training the TransVert+ model requires radiograph images and their corresponding 'real world' standing 3D spine models. However, this correspondence does not exist. It is, in effect, the issue we aim to address. Thus, TransVert+ is trained on sagittal and coronal digitally reconstructed radiographs (DRR) generated from prone CT images with supervision from the voxel-level, vertebral segmentation masks of the corresponding CT images. Since DRRs are similar in appearance to real radiographs, one can deploy a DRR-trained TransVert+ model on clinical standing radiographs.

Generating 3D shapes from 2D information is an ill-posed problem. We model this task as a registration problem. Given 2D orthogonal information we estimate vector fields to deform shape templates to match the target 3D shape. We introduce a novel deep neural architecture to fuse the 2D information from orthogonal views and estimate the vector field required to maximize the alignment of the deformed template and the target shape. More specifically, to infer 3D spinal shapes, given 2D orthogonal radiographs and vertebral centroid coordinates in the global coordinate system, our model predicts a vector field to deform each vertebral shape template to match the target. The vector fields are predicted by considering each vertebral shape locally and also considering the spinal curvature globally. We show that incorporating the global spinal shape information improves the performance of the model. Our model is composed of two

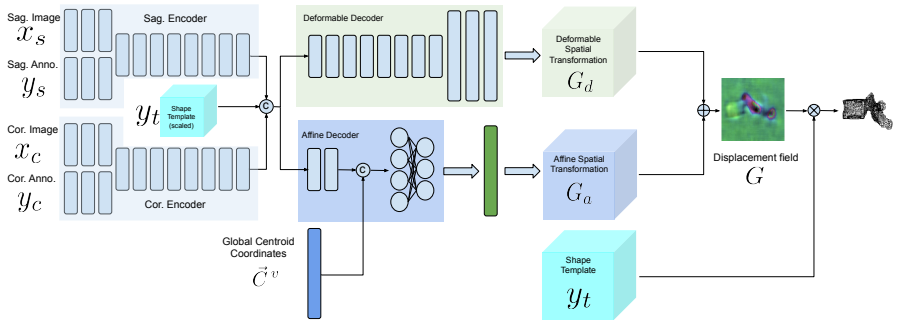


Figure A.3: Architecture of *TransVert+*. Our model is composed of sagittal and coronal 2D encoders, an affine 3D decoder and a deformable 3D decoder. The down-scaled shape templates are concatenated with the features extracted by encoders and fed to the decoders. The centroid coordinates in the global spine coordinate system are concatenated with the affine decoder feature maps. The $\otimes$, $\oplus$ and $\|$ operators represent warping, addition and concatenation respectively.

encoders for each view, an affine decoder and an deformable decoder. In the following sections we describe the model architecture and training scheme in detail.

A.3.2 Network Architecture

Fig. A.3 demonstrates a block diagram of the model and its subnetworks. The model is composed of two encoders, one for each view (a sagittal encoder and a coronal encoder) and two decoders (an affine decoder and a deformable decoder). The input to the encoders is the image patch of the vertebra from the radiograph and the centroid annotation on the vertebra of interest. The features extracted using the encoders are concatenated to the global centroid coordinates and fed to the affine decoder (to estimate the affine transformation parameters for each vertebra) and the deformable decoder (which is a fully convolutional network to estimate the deformation for each voxel). Finally the affine and deformable vector fields are summed to produce the final displacement field, which is used to warp the template and produce the 3D shape model.

A.3.2.1 Sagittal and Coronal Encoders

The architecture of sagittal and coronal encoders are designed differently. In Fig. A.4 the architecture for each view is visualized. Each encoder is designed to reconstruct the missing third dimension of the 2D input. Therefore, they contain anisotropic convolutions with the longer side along the dimension that needs to be expanded. For instance, for a coronal input the anterior-posterior dimension needs to be expanded. For the same reason, the convolutional strides and padding directions are orthogonal for each of the views. We empirically observed that, employing 'squeeze and excitation' block in the network results in a better performance than a naive fusion by concatenating the multiple channels. As input to the encoders, the vertebral images and VOI-annotations are

Figure A.4: Architecture of the orthogonal encoders, which employ anisotropic convolutions, with an anisotropy along the dimensions that need to be expanded. We use 'squeeze and excitation' blocks (depicted in red) to fuse the image features and annotation features.

combined using a 'squeeze and excitation' block (depicted in red)[127]. In our previous work [124] we showed that, using encoders for each view with anisotropic convolutions, outperforms using a single encoder or fusing input 2D data by outer product of the orthogonal 2D images.

A.3.2.2 Affine Decoder

Since the vertebrae are not completely visible in the radiographs, generating the 3D spine model given only the encoded features for each vertebra separately could lead to inaccurate orientation.

Including the vertebral centroid coordinates in the global coordinate system provides the model with a holistic shape of the spine. For example, Fig. A.1 demonstrates lateral and anterior-posterior (a.p.) view radiographs of three different patients. The distance between the centroids determines the scale. Also, if one fits a curve to all of the vertebral centroids, the orientation of each vertebra should be almost perpendicular to the curve at each vertebral centroid. Although defining the orientation of the vertebrae and the other constraints like inter-vertebral distance need accurate vertebral landmark detection on radiographs, our experiments show that we can train an Multi Layer Perceptron (MLP) to estimate the 3D affine parameters, given only the vertebral centroids and the information extracted by the encoders.

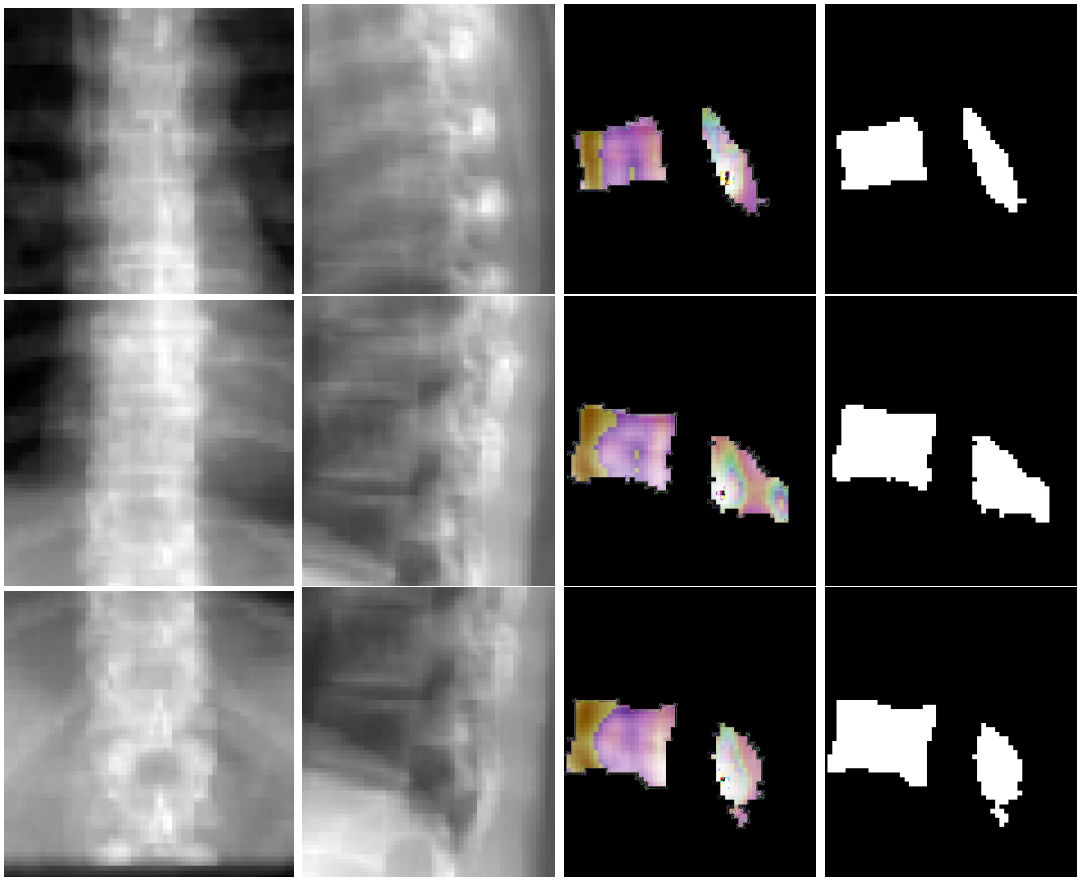


Figure A.5: Visualization of coronal and sagittal image patches from three vertebrae. First and second columns are the coronal and sagittal image patches, third column shows one slice of the predicted deformation field and last column is the corresponding slice in the resulting shape.

Since the features calculated by the encoders for data samples in a batch are independent of each other, we need an intra-batch fusion mechanism to give the model a holistic view of the features from different spinal regions. We assume that the vertebrae in a batch are from the same spine and in order (we train the model in this way). For intra-batch fusion, we flatten the features from both encoders and also the down-scaled vertebral shape templates and concatenate them with the vertebral centroid coordinates, for all vertebrae. These inputs are fed into our affine subnetwork, which is a MLP.

For each vertebra, the affine subnetwork estimates four parameters: the scaling factor (S), and the rotation about each axis (θ_x, θ_y, θ_z). After the transformation matrix is created, the corresponding affine vector field is calculated. We do not include translation, since the input image patches are extracted in such a way that the vertebral centroid is at a fixed location in the image patch.

$$\theta_x, \theta_y, \theta_z, S = AffineDecoder(x_s, x_c, y_s, y_c, \vec{C}^{\,v}), \tag{A.5}$$

$$T_{affine} = R(\theta_x) * R(\theta_y) * R(\theta_z) * T(S), \tag{A.6}$$

$$R(\theta_x) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) & 0 \\ 0 & \sin(\theta_x) & \cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.7}$$

$$R(\theta_y) = \begin{bmatrix} \cos(\theta_y) & 0 & \sin(\theta_y) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta_y) & 0 & \cos(\theta_y) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.8}$$

$$R(\theta_z) = \begin{bmatrix} \cos(\theta_z) & -\sin(\theta_z) & 0 & 0 \\ \sin(\theta_z) & \cos(\theta_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.9}$$

$$T(S) = \begin{bmatrix} S & 0 & 0 & 0 \\ 0 & S & 0 & 0 \\ 0 & 0 & S & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.10}$$

$$Affine_{vf} = \tau(T_{affine}), \tag{A.11}$$

where the $Affine_{vf}$ is the affine vector field and $\tau()$ is the function to generate an affine vector field given the transformation matrix.

A.3.2.3 Deformable Decoder

In parallel with affine parameter estimation for rough alignment between the template and target shapes, we synthesize the finer details. We design a subnetwork termed the "deformable decoder" for estimating local shape deformations. The deformable decoder is a fully convolutional network. The inputs to this network are the intermediate 3D latent representation calculated by the encoders with the down-scaled (the same size as 3D latent representation) shape template as an extra channel. Including the shape template as another input channel helps the model to infer the differences between the template and target shapes. The deformable decoders's output is a 3D vector field with target volume resolution. Contrary to the affine decoder, which estimates the affine transformation parameters based on the entire batch, the deformable decoder is focused on single data samples in the batch, to consider the finer shape details for each sample,

$$Deformable_{vf} = DeformableDecoder(x_s, x_c, y_s, y_c). \tag{A.12}$$

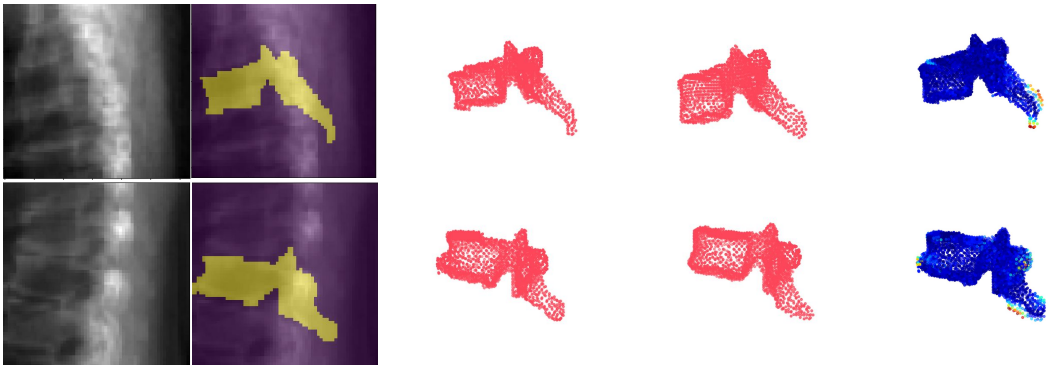


Figure A.6: Shape modelling with TransVert+ on DRRs: The first column indicates the image input. The second and third columns visualize the ground truth (GT) vertebral mask, and the fourth visualizes the predicted 3D shape model. The last column shows an overlayed Chamfer distance map between point clouds of GT and prediction.

A.3.3 Learning

We train the model using the ℓ_1 distance between the deformed templates and the target vertebral shapes in the voxel space. To regularize the deformations, we incorporate a smoothing term in the cost function. Solely using a regression loss leads to convergence to a local optimum in which a mean (or median) shape is predicted, especially in the highly varying regions of the vertebra such as the vertebral processes. We have,

$$\mathcal{L}_{total} = \alpha_p \mathcal{L}_{\ell_1} + \alpha_s \mathcal{L}_{smooth} \tag{A.13}$$

$$\mathcal{L}_{\ell_1} = ||\mathbf{y} - G(x_s, x_c, y_s, y_c, y_t, \vec{C}^v) \circ y_t||_1 \tag{A.14}$$

$$\mathcal{L}_{smooth} = \frac{1}{X.Y.Z} \int_0^X \int_0^Y \int_0^Z ||G(x_s, x_c, y_s, y_c, y_t, \vec{C}^v)||_1 dxdydz, \tag{A.15}$$

where α_p and α_s are weights for the network prediction and smoothness terms respectively, and are fixed to $\alpha_p = 10$ and $\alpha_s = 0.1$. Note that $\mathbf{y}$ contains an integer value of $\{0, i\}$, where $i \in \{8, 9, \dot{2}4\}$ represents the vertebral index from T1 to L5. Constraining the network to predict the vertebral index implicitly requires it to learn relating the vertebral index to the shape as a prior.

In order to impose a constraint for locally smooth deformations and a minimum displacement solution for our registration problem, we add $\mathcal{L}_{smooth}$ to penalize the $\ell_1 - norm$ of the deformation field.

The network was implemented with the Pytorch framework on a Quadro P6000 GPU. It was trained until convergence using the Adam optimizer [128] with an initial learning rate of 0.0001.

Setup	Dice	Hausdorff (mm)	nWESD
AffTransVert	0.8847	12.40	0.3181
DefTransVert	0.9405	9.08	0.2430
TransVert	0.9426	7.93	0.2779
TransVert+	0.9510	5.70	0.1797

Table A.1: Architectural ablation study: The performance progressively improves with addition of each component. While TransVert outperforms the separate Affine and Deformable TransVert models, including both of the Affine and the Deformable decoders in TrasVert+ model performs better than TransVert.

Setup	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	L1	L2	L3	L4
Aff	0.89	0.87	0.87	0.88	0.87	0.87	0.87	0.87	0.88	0.89	0.90	0.90	0.90	0.88	0.87	0.85
Def	0.93	0.94	0.93	0.93	0.93	0.95	0.95	0.96	0.96	0.97	0.96	0.95	0.94	0.93	0.90	0.91
TVert	0.94	0.93	0.94	0.94	0.95	0.95	0.95	0.96	0.96	0.97	0.96	0.95	0.94	0.93	0.91	0.91
TVet+	0.95	0.93	0.94	0.95	0.95	0.95	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.95	0.92	0.92

Table A.2: Vertebra-wise comparison of different architectures (Affine, Deformable, TransVert and TransVert+) using Dice scores. A higher score indicates better performance.

A.4 Validation

Training the TransVert+ model requires radiograph images and their corresponding 3D spine shapes. However, this correspondence does not exist. Therefore, we train TransVert+ on sagittal and coronal DRRs generated of CT images and supervised by the corresponding CT images' vertebral segmentation masks. In this section, we describe the process of generating DRRs, report an ablative study on architectural choices, and evaluate TransVert+'s performance on clinical radiographs.

A.4.1 Data

Recall that TransVert+ works with two data modalities: it is trained on DRRs extracted from CT images and is deployed on clinical radiographs.

A.4.1.1 CT data

We employed two sets of data: First, a public dataset for lung nodule detection with 800 chest CT scans [97], and second, an in-house dataset with 154 spinal CT scans. Overall, we work with $\sim 12K$ vertebrae split $5:1$ forming the training and validation set and report 5-fold cross-validated results. Of note, [97] is a lung-centred dataset, thus consisting of few lumbar vertebrae. All of spinal CT scans were resampled to $1mm$ resolution and segmented using [4]. Next, an experienced neuro-radiologist approved the generated masks to consider only accurate ones for the study. Consequently, we excluded 50 cases from the datset.

We employ a ray-casting approach [98] to construct DRRs from CT scan. In this method, we define lines from the radiation source (focal point) to every single pixel on

Setup	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	L1	L2	L3	L4	L5
Aff	12.6	10.5	10.8	11.3	11.3	11.3	11.6	11.6	11.9	12.2	12.7	13.0	13.1	13.4	13.8	14.8	14.5
Def	9.1	8.4	8.5	8.4	8.7	8.5	8.6	8.7	8.9	8.8	9.2	8.9	9.0	9.5	10.1	10.2	10.5
TVert	7.4	7.2	7.3	7.2	7.9	7.5	7.4	7.5	7.5	7.6	8.0	7.8	7.8	8.4	8.9	9.3	9.1
TVert+	5.6	5.0	5.0	4.9	5.1	5.0	5.1	5.2	5.2	5.3	5.5	5.4	5.4	5.9	6.5	8.4	8.5

Table A.3: Vertebra-wise comparison of different architectures using Hausdorff distance. A lower distance indicates better performance.

Setup	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	L1	L2	L3	L4	L5
Aff	0.31	0.29	0.33	0.26	0.28	0.34	0.32	0.32	0.34	0.32	0.31	0.29	0.28	0.29	0.33	0.38	0.37
Def	0.21	0.21	0.25	0.24	0.28	0.24	0.25	0.27	0.25	0.23	0.20	0.26	0.30	0.16	0.24	0.35	0.36
TVert	0.24	0.24	0.24	0.23	0.29	0.27	0.27	0.26	0.24	0.22	0.22	0.24	0.28	0.27	0.36	0.38	0.40
TVert+	0.15	0.13	0.16	0.10	0.19	0.18	0.20	0.15	0.12	0.22	0.19	0.11	0.15	0.15	0.29	0.31	0.30

Table A.4: Vertebra-wise comparison of different architectures using nWESD distance. A lower distance indicates better performance.

the DRR image and calculate the integral of the CT intensities over these lines. In this simulation we assign (180cm) and (150cm) to the radiation source-to-detector distance and the source-to-object distance parameters respectively. An example of orthogonal DRRs generated from a CT scan is shown in Fig. A.7. Once the sagittal and coronal DRRs are generated, the inputs for TransVert+ are constructed by extracting image patches of size 64×64 around each vertebral centroid. Similarly, the VOI-annotations were also extracted automatically from the projected segmentation mask.

A.4.1.2 Clinical radiographs

We validate TransVert+ on clinical, standing radiographs (pairs of lateral and anterior-posterior projections) aquired from 30 patients. Before deploying our TransVert+ model on the clinical radiographs, we resample all radiographs to $1mm$ resolution. Image acquisition parameters such as the source-to-detector and source-to-object distances were similar to those used for DRR generation. We employed [99, 129] to automatically generate the vertebral annotations on both views.

A.4.1.3 Image normalization

We trained TransVert+ on DRRs and tested it on real clinical radiographs. To enable a transfer of learning between between these modalities, we need to normalize the intensities to a similar range. Therefore, we employ z-score normalization, i.e. $\mathcal{I} = {}^{(\mathcal{I}-\mu_{\mathcal{I}})}/_{/\sigma_{\mathcal{I}}}$, where $\mu_{\mathcal{I}}$ and $\sigma_{\mathcal{I}}$ are the mean and standard deviation of the image $\mathcal{I}$, respectively.

A.4.2 Metrics

The Dice score is the most popular measurement for evaluating segmentation accuracy and measures the overlap between two binary images. However, the Dice score is a poor measure of segmentation accuracy when the shapes to be compared are not "blob-like".

Another popular metric for segmentation evaluation is the Hausdorff distance. The Hausdorff distance between two segmentations represented as surface meshes is the maximum distance from vertices on the first mesh to the vertices on the second mesh.

The measurements like the Dice score, average many local errors measured at each voxel. Thus, a segmentation that is largely correct with a few major shape errors will have the same score as a segmentation that is only slightly wrong everywhere.

As mentioned before, we do not have the 3D ground truth for the clinical radiograph of a patient. We can acquire the 3D CT segmentation of the same patient. However, because of difference in spinal posture the vertebral orientations are different and we cannot compare the 3D vertebrae reconstructed from clinical radiographs to the ones from CT using Dice score or Hausdorff distance. Thus, we desire a rotation-invariant metric to evaluate the model performance on clinical radiographs, for each individual vertebra.

The normalized weighted spectral distance (nWESD) [130, 131] is a global shape measure based on heat trace analysis via the Laplace operator. The eigenvalues of the Laplacian of a shape are strongly connected to the shape's geometric properties, such as its volume, surface area and mean curvature. The Laplace spectrum is invariant to isometric transformations (rigid transformations), and changes continuously as a shape's boundary is transformed. Thus, because of rotational invariance, we can use this metric for comparing the 3D vertebral shapes reconstructed from the patient's clinical radiographs to the vertebral shapes from his 3D CT segmentations.

The weighted spectral distance (WESD) between two binary segmentations Ω_λ and Ω_ξ is defined as

$$\rho(\Omega_\lambda, \Omega_\xi) = [\sum_{n=1}^{\infty} (\frac{|\lambda_n - \xi_n|}{\lambda_n \xi_n})^p]^{1/p}, \tag{A.16}$$

where d indicates the dimensionality of the binary segmentations, λ_n and ξ_n denote eigenvalues of the segmentations Ω_λ and Ω_ξ, and $p \in \mathbb{R}$ with $p > d/2$.

The normalized WESD (nWESD) is derived using the fact that WESD converges as $N \to \infty$ (even though each eigenvalue spectrum is divergent) and is bounded above by $W(\Omega_\lambda, \Omega_\xi)$ (for details, see [130]). Therefore, the (finite) nWESD $\bar{\rho}^N(\Omega_\lambda, \Omega_\xi)$ can be defined as:

$$\bar{\rho}^N(\Omega_\lambda, \Omega_\xi) = \frac{\rho^N(\Omega_\lambda, \Omega_\xi)}{W(\Omega_\lambda, \Omega_\xi)} \in [0, 1) \tag{A.17}$$

$$\rho^N(\Omega_\lambda, \Omega_\xi) = [\sum_{n=1}^{N} (\frac{|\lambda_n - \xi_n|}{\lambda_n \xi_n})^p]^{1/p}, \tag{A.18}$$

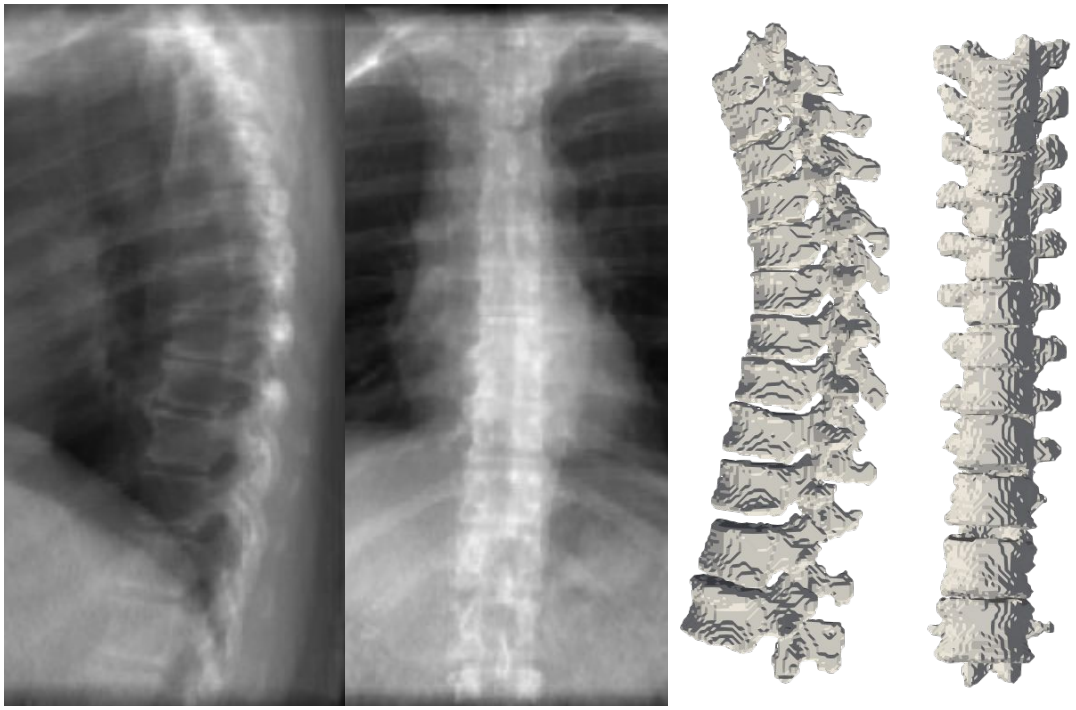


Figure A.7: Orthogonal DRRs of a patient and reconstructed 3D spine model. DRRs are generated to from CT scans and our model is trained to reconstruct the 3D spine shape from DRRs.

A.4.3 Experiments

In order to analyze the contribution of various architectural components of the TransVert+ and to validate it's performance on clinical radiographs, we propose three sets of experiments. The performance evaluation in various settings was compared by computing the Dice coefficient, Hausdorff distance and normalized weighted spectral distance (nWESD) between the predicted 3D vertebral mask and the ground truth CT mask, where appropriate.

A.4.3.1 Analysing TransVert+'s architecture

The proposed architecture for TransVert+ consists of the following architectural choices: fusion of sagittal and coronal views with shape prior, the affine decoder and a convolutional decoder to estimate the displacement field for each vertebrae. We performed an ablative study over these components. First, we evaluated the performance of the model working only with the affine decoder. Second, we trained the model solely with the deformable decoder to estimate the deformation fields. Third, we repeat the performance our previous model TransVert, in which we did not incorporate the shape priors,

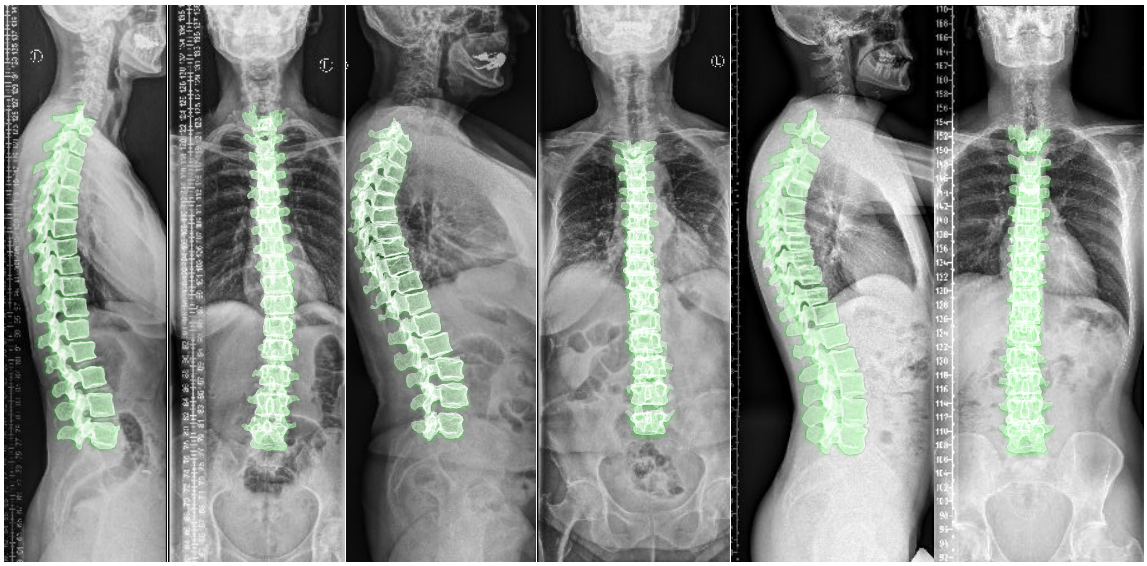


Figure A.8: Full 3D spine models: 3D patient-specific spine models constructed from real clinical radiographs. Each sagittal and coronal view radiograph pair is from a different patient.

where we estimate the vertebral shapes directly in the voxel space without registration. Firnally, we give results for TransVert+.

The results are reported in Table A.1. As we expected, the affine results are not better that TransVert, since the model roughly aligns the shape template to the target and the results lack shape details. Training the model with a deformable decoder performs better than the affine model. The deformable model achieves an average Dice score of 0.9405, close to that of TransVert model (0.9426). In the Hausdorff distance metric, TransVert performs better but for the nWESD metric, the defromable model achieves a better score. Finally, our main TransVert+ model which incorporates the global shape information of spinal curvature using the affine decoder and takes care of the local shape details using the deformable decoder, outperformed all of the previous models in all metrics. For a detailed comparison of the methods, in Tables A.2, A.3 and A.4 we report the mean Dice score, Hausdorff distance and nWESD distance for each vertebra. Fig. A.5 depicts coronal and sagittal image patches from three vertebrae, mid-slice of the resulting estimated displacement field and vertebral shape. Fig. A.6 shows an example point cloud (with 2048 points) from the predicted and ground truth shapes, along with a point-wise Chamfer distance map. Observe that the vertebra's posterior region (vertebral process) is hardly visible in the image inputs. In spite of this, TransVert+ was able to reconstruct the 3D shape of vertebral process.

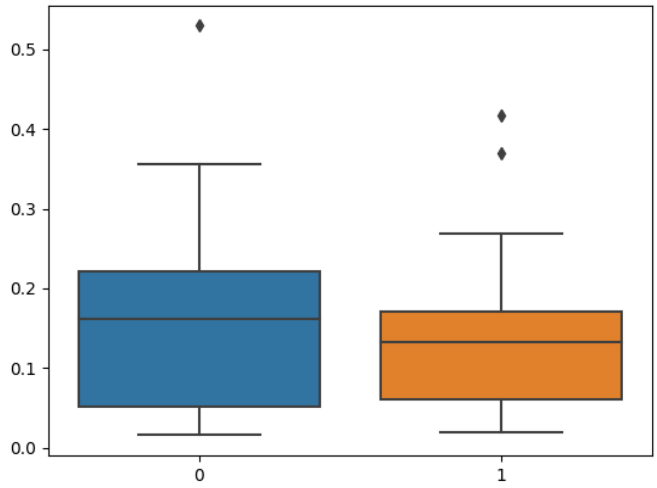


Figure A.9: Comparing vertebral shapes to the ones predicted from radiographs using TransVert (Blue) and TransVert+ (Orange) using the nWESD metric. The mean nWESD metric for TransVert+ is lower than TransVert, indicating a better performance.

A.4.3.2 2D-to-3D translation in clinical radiographs

TransVert+ generates 3D shapes of each vertebra. To reconstruct the 3D shape of the full spine, we stack the predicted vertebral shapes at their corresponding 3D centroid locations. We obtain the sagittal and coronal centroid coordinates elements from sagittal and coronal reformations, respectively. Fig. A.8 depicts the results of deploying TransVert+ for reconstructing the 3D, patient-specific posture of the upright standing spine. As stated, there is no 3D ground truth spinal model for the clinical radiographs. Observe the matched reconstruction of the 3D spine posture to the spinal posture in radiographs.

A.4.3.3 Quantitative evaluation of performance on clinical radiographs

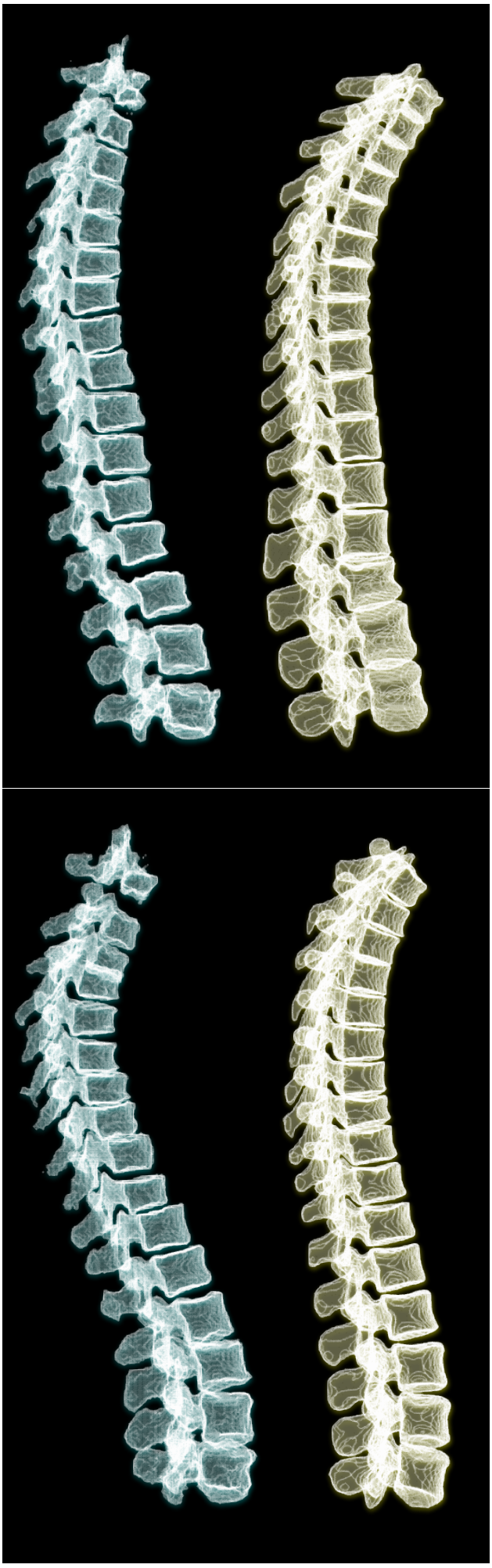
We quantitatively evaluated the performance of our model on clinical radiographs in two patients for whom we have CT scans in addition to orthogonal standing radiographs. Using our TransVert model and the TransVert+ model proposed in this work, we generated the 3D shapes of the vertebrae given the 2D clinical radiographs. Then we compared each vertebra to the corresponding one from the CT scan segmentation masks, which refer to the same object but in a different orientation. Conventional metrics like the Dice score or Hausdorff are not applicable in this case, but we can use the nWESD metric as it is invariant to rigid transformations (including rotation). The resulting nWESD scores are demonstrated in Fig. A.9, where the mean nWESD was

0.13 for TransVert+. To verify that the performance improvements of TransVert+ over TransVert on clinical radiographs were statistically significant, we conducted an ANOVA analysis on the nWESD scores, yielding $p - value = 0.001$ and showing the statistically significant improvement.

To appreciate the difference of spine posture in the standing and lying down positions, we illustrate the two cases we used in this experiment in Fig. A.10. The spine postures reconstructed from radiographs in upright standing position are depicted in left and the ones from CT are depicted in right side of the figure.

A.5 Conclusion

We introduced TransVert+, a neural network architecture to reconstruct a full 3D spinal model from 2D orthogonal radiographs by deforming vertebral shapes. We proposed a supervised approach to train the model on synthetic data (DRRs) and transfer the trained model to radiographs. We demonstrated an improved performance of the model after incorporating shape priors into the model and separating the registration problem into affine and deformable registrations. We improved upon the state-of-the-art in three different metrics for this task. Finally, we successfully validated our model on real-world clinical radiographs and quantitatively compared the results to CT segmentations of the same patient for the first time. The approach we proposed could be used to simulate the spinal posture in standing position under weight bearing, and used in bio-mechanical analysis.



Acknowledgment

This work is supported by the European Research Council (ERC) under the European Union's 'Horizon 2020' research & innovation programme (GA637164–iBack–ERC–2014–STG). The authors would like to thank Dr. Polina Golland and Dr. Justin Solomon. We also acknowledge support of NIH NIBIB NAC P41EB015902 and Philips Inc.

Bibliography

[1] E. Melhem, A. Assi, R. El Rachkidi, and I. Ghanem. Eos® biplanar x-ray imaging: concept, developments, benefits, and limitations. *Journal of children's orthopaedics*, 10(1):1–14, 2016.

[2] A. Bayat, A. Sekuboyina, J. C. Paetzold, C. Payer, D. Stern, M. Urschler, J. S. Kirschke, and B. H. Menze. Inferring the 3d standing spine posture from 2d radiographs. *arXiv preprint arXiv:2007.06612*, 2020.

[3] J. Radon. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 5:21, 2005.

[4] A. Sekuboyina, A. Bayat, M. E. Husseini, M. Löffler, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinitsch, C. Payer, M. Urschler, et al. Verse: A vertebrae labelling and segmentation benchmark. *arXiv preprint arXiv:2001.09193*, 2020.

[5] J.-T. Lu, S. Pedemonte, B. Bizzo, S. Doyle, K. P. Andriole, M. H. Michalski, R. G. Gonzalez, and S. R. Pomerantz. Deep spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In *Machine Learning for Healthcare Conference*, pages 403–419. PMLR, 2018.

[6] M. Löffler, N. Sollmann, K. Mei, A. Valentinitsch, P. Noël, J. Kirschke, and T. Baum. X-ray-based quantitative osteoporosis imaging at the spine. *Osteoporosis International*, 31(2):233–250, 2020.

[7] D. P. Anitha, T. Baum, J. S. Kirschke, and K. Subburaj. Effect of the intervertebral disc on vertebral bone strength prediction: a finite-element study. *The Spine Journal*, 20(4):665–671, 2020.

[8] F. Laouissat, A. Sebaaly, M. Gehrchen, and P. Roussouly. Classification of normal sagittal spine alignment: refounding the roussouly classification. *European Spine Journal*, 27(8):2002–2011, 2018.

[9] T. R. Oxland. Fundamental biomechanics of the spine—what we have learned in the past 25 years and future directions. *Journal of biomechanics*, 49(6):817–832, 2016.

[10] G. von Campe and K. Pistracher. Patient specific implants (psi). In *Cranial Implant Design Challenge*, pages 1–9. Springer, 2020.

[11] T. Zegers, M. ter Laak-Poort, D. Koper, B. Lethaus, and P. Kessler. The therapeutic effect of patient-specific implants in cranioplasty. *Journal of Cranio-Maxillofacial Surgery*, 45(1):82–86, 2017.

[12] J. Li, A. Pepe, C. Gsaxner, and J. Egger. An online platform for automatic skull defect restoration and cranial implant design. In *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11598, page 115981Q. International Society for Optics and Photonics, 2021.

[13] X. Chen, L. Xu, X. Li, and J. Egger. Computer-aided implant design for the restoration of cranial defects. *Scientific reports*, 7(1):1–10, 2017.

[14] F. P. Oliveira and J. M. R. Tavares. Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014.

[15] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, 2020.

[16] R. Shams, P. Sadeghi, R. A. Kennedy, and R. I. Hartley. A survey of medical image registration on multicore and the gpu. *IEEE signal processing magazine*, 27(2):50–60, 2010.

[17] P. Zhilkin and M. Alexander. 3d image registration using a fast noniterative algorithm. *Magnetic Resonance Imaging*, 18(9):1143–1150, 2000.

[18] O. Zvitia, A. Mayer, R. Shadmi, S. Miron, and H. K. Greenspan. Co-registration of white matter tractographies by adaptive-mean-shift and gaussian mixture modeling. *IEEE Transactions on Medical Imaging*, 29(1):132–145, 2009.

[19] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes, and S. Ourselin. A registration-based propagation framework for automatic whole heart segmentation of cardiac mri. *IEEE transactions on medical imaging*, 29(9):1612–1625, 2010.

[20] W. Bai and M. Brady. Motion correction and attenuation correction for respiratory gated pet images. *IEEE transactions on medical imaging*, 30(2):351–365, 2010.

[21] S. K. Balci, P. Golland, and W. Wells. Non-rigid groupwise registration using b-spline deformation model. *Open source and open data for MICCAI*, pages 105–121, 2007.

[22] M. Khader and A. B. Hamza. An entropy-based technique for nonrigid medical image alignment. In *International Workshop on Combinatorial Image Analysis*, pages 444–455. Springer, 2011.

[23] M. C. H. Lee, K. Petersen, N. Pawlowski, B. Glocker, and M. Schaap. Tetris: template transformer networks for image segmentation with shape priors. *IEEE transactions on medical imaging*, 38(11):2596–2606, 2019.

[24] M. Auer, P. Regitnig, and G. A. Holzapfel. An automatic nonrigid registration for stained histological sections. *IEEE Transactions on Image Processing*, 14(4):475–486, 2005.

[25] C. R. Meyer, J. L. Boes, B. Kim, P. H. Bland, G. L. Lecarpentier, J. B. Fowlkes, M. A. Roubidoux, and P. L. Carson. Semiautomatic registration of volumetric ultrasound scans. *Ultrasound in medicine & biology*, 25(3):339–347, 1999.

[26] M. Holden. A review of geometric transformations for nonrigid body registration. *IEEE transactions on medical imaging*, 27(1):111–128, 2007.

[27] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[28] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019.

[29] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[30] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[31] A. Sedghi, L. J. O'Donnell, T. Kapur, E. Learned-Miller, P. Mousavi, and W. M. Wells III. Image registration: Maximum likelihood, minimum entropy and deep learning. *Medical Image Analysis*, 69:101939, 2021.

[32] X. Cheng, L. Zhang, and Y. Zheng. Deep similarity learning for multimodal medical images. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):248–252, 2018.

[33] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P. A. Pinto, B. J. Wood, and P. Yan. Learning deep similarity metric for 3d mr–trus image registration. *International journal of computer assisted radiology and surgery*, 14(3):417–425, 2019.

[34] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention*, pages 10–18. Springer, 2016.

[35] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[36] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. An information theoretic approach for non-rigid image registration using voxel class probabilities. *Medical image analysis*, 10(3):413–431, 2006.

[37] W. Bai, W. Shi, D. P. O'regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images. *IEEE transactions on medical imaging*, 32(7):1302–1315, 2013.

[38] K. A. Saddi, C. Chefd'Hotel, M. Rousson, and F. Cheriet. Region-based segmentation via non-rigid template matching. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE, 2007.

[39] S. Liu, I. Ago, and C. L. Giles. Learning a hierarchical latent-variable model of voxelized 3d shapes. *arXiv preprint arXiv:1705.05994*, 2017.

[40] A. Sharma, O. Grau, and M. Fritz. Vconv-dae: Deep volumetric shape learning without object labels. In *European Conference on Computer Vision*, pages 236–250. Springer, 2016.

[41] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pages 484–499. Springer, 2016.

[42] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. W. Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

[43] D. Stutz and A. Geiger. Learning 3d shape completion under weak supervision. *International Journal of Computer Vision*, 128(5):1162–1181, 2020.

[44] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin, and D. Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2484–2493, 2015.

[45] C. Hane, N. Savinov, and M. Pollefeys. Class specific 3d object shape priors using surface normals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–659, 2014.

[46] F. Engelmann, J. Stückler, and B. Leibe. Joint object pose estimation and shape reconstruction in urban street scenes using 3d shape priors. In *German Conference on Pattern Recognition*, pages 219–230. Springer, 2016.

[47] D. T. Nguyen, B.-S. Hua, K. Tran, Q.-H. Pham, and S.-K. Yeung. A field model for repairing 3d shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5684, 2016.

[48] D. Stutz and A. Geiger. Learning 3d shape completion under weak supervision. *International Journal of Computer Vision*, pages 1–20, 2018.

[49] G. Riegler, A. O. Ulusoy, H. Bischof, and A. Geiger. Octnetfusion: Learning depth fusion from data. In *2017 International Conference on 3D Vision (3DV)*, pages 57–66. IEEE, 2017.

[50] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[51] M. Koziński, A. Mosinska, M. Salzmann, and P. Fua. Learning to segment 3d linear structures using only 2d annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–291. Springer, 2018.

[52] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[53] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *arXiv preprint arXiv:1610.07584*, 2016.

[54] C. Häne, S. Tulsiani, and J. Malik. Hierarchical surface prediction for 3d object reconstruction. In *2017 International Conference on 3D Vision (3DV)*, pages 412–420. IEEE, 2017.

[55] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *arXiv preprint arXiv:1612.00814*, 2016.

[56] S. Tulsiani, A. A. Efros, and J. Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2897–2905, 2018.

[57] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018.

[58] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[59] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017.

[60] J. Egger, J. Li, X. Chen, U. Schäfer, G. von Campe, M. Krall, U. Zefferer, C. Gsaxner, A. Pepe, and D. Schmalstieg. Towards the automatization of cranial implant design in cranioplasty. *Zenodo. http://doi.org/10.5281/zenodo.3715953*, 2020.

[61] O. Ronneberger, P. Fischer, and T. 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.

[62] A. Bayat, S. Shit, A. Kilian, J. T. Liechtenstein, J. S. Kirschke, and B. H. Menze. Cranial implant prediction using low-resolution 3d shape completion and high-resolution 2d refinement. In *Cranial Implant Design Challenge*, pages 77–84. Springer, 2020.

[63] M. Dreischarf, A. Shirazi-Adl, N. Arjmand, A. Rohlmann, and H. Schmidt. Estimation of loads on human lumbar spine: a review of in vivo and computational model studies. *Journal of biomechanics*, 49(6):833–845, 2016.

[64] M. Akhavanfar, H. Kazemi, A. Eskandari, and N. Arjmand. Obesity and spinal loads; a combined mr imaging and subject-specific modeling investigation. *Journal of biomechanics*, 70:102–112, 2018.

[65] Z. El Ouaaid, A. Shirazi-Adl, and A. Plamondon. Effect of changes in orientation and position of external loads on trunk muscle activity and kinematics in upright standing. *Journal of Electromyography and Kinesiology*, 24(3):387–393, 2014.

[66] A. Eskandari, N. Arjmand, A. Shirazi-Adl, and F. Farahmand. Subject-specific 2d/3d image registration and kinematics-driven musculoskeletal model of the spine. *Journal of Biomechanics*, 57:18–26, 2017.

[67] S. Bauer, U. Hausen, and K. Gruber. Effects of individual spine curvatures–a comparative study with the help of computer modelling. *Biomedical Engineering/Biomedizinische Technik*, 57(SI-1 Track-O):132–135, 2012.

[68] M. Ketcha, T. De Silva, A. Uneri, M. Jacobson, J. Goerres, G. Kleinszig, S. Vogt, J. Wolinsky, and J. Siewerdsen. Multi-stage 3d–2d registration for correction of anatomical deformation in image-guided spine surgery. *Physics in Medicine & Biology*, 62(11):4604, 2017.

[69] L. Humbert, J. A. De Guise, B. Aubert, B. Godbout, and W. Skalli. 3d reconstruction of the spine from biplanar x-rays using parametric models based on transversal and longitudinal inferences. *Medical engineering & physics*, 31(6):681–687, 2009.

[70] S. Kadoury, F. Cheriet, and H. Labelle. Personalized x-ray 3-d reconstruction of the scoliotic spine from hybrid statistical and image-based models. *IEEE Transactions on medical imaging*, 28(9):1422–1435, 2009.

[71] B. Aubert, C. Vazquez, T. Cresson, S. Parent, and J. A. de Guise. Toward automated 3d spine reconstruction from biplanar radiographs using cnn for statistical spine model fitting. *IEEE transactions on medical imaging*, 38(12):2796–2806, 2019.

[72] X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10619–10628, 2019.

[73] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2690–2698, 2019.

[74] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun. Pix2vox++: multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020.

[75] F. Milletari, A. Rothberg, J. Jia, and M. Sofka. Integrating statistical prior knowledge into convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 161–168. Springer, 2017.

[76] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017.

[77] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[78] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz. Automated model-based vertebra detection, identification, and segmentation in ct images. *Medical image analysis*, 13(3):471–482, 2009.

[79] H. Mirzaalian, M. Wels, T. Heimann, B. M. Kelm, and M. 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.

[80] S. Ruiz-España, J. Domingo, A. Díaz-Parra, E. Dura, V. D'Ocón-Alcañiz, E. Arana, and D. Moratal. Automatic segmentation of the spine by means of a probabilistic atlas with a special focus on ribs suppression. preliminary results. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2014–2017. IEEE, 2015.

[81] S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, and C. 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.

[82] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. 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.

[83] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. 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.

[84] Y. Zhan, D. Maneesh, M. Harder, and X. S. 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.

[85] Y. Cai, S. Osman, M. Sharma, M. Landis, and S. Li. Multi-modality vertebra recognition in arbitrary views using 3d deformable hierarchical model. *IEEE transactions on medical imaging*, 34(8):1676–1693, 2015.

[86] S. Caprara, F. Carrillo, J. G. Snedeker, M. Farshad, and M. Senteler. Automated pipeline to generate anatomically accurate patient-specific biomechanical models of healthy and pathological fsus. *Frontiers in Bioengineering and Biotechnology*, 9, 2021.

[87] A. Sekuboyina, A. Bayat, M. E. Husseini, M. Löffler, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinitsch, C. Payer, M. Urschler, et al. Verse: A vertebrae labelling and segmentation benchmark. *arXiv preprint arXiv:2001.09193*, 2020.

[88] P. A. Fishwick. Simpack: getting started with simulation programming in c and c++. In *Proceedings of the 24th conference on Winter simulation*, pages 154–162, 1992.

[89] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using boostnet. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 127–135. Springer, 2017.

[90] D. Yang, T. Xiong, D. Xu, Q. Huang, D. Liu, S. K. Zhou, Z. Xu, J. Park, M. Chen, T. 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.

[91] D. Yang, T. Xiong, D. Xu, S. K. Zhou, Z. Xu, M. Chen, J. Park, S. Grbic, T. D. Tran, S. P. Chin, et al. Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3d ct volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 498–506. Springer, 2017.

[92] C. Payer, D. Štern, H. Bischof, and M. Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical image analysis*, 54:207–219, 2019.

[93] B. Glocker, J. Feulner, A. Criminisi, D. R. Haynor, and E. 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.

[94] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. 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.

[95] A. Suzani, A. Seitel, Y. Liu, S. Fels, R. N. Rohling, and P. Abolmaesumi. Fast automatic vertebrae detection and localization in pathological ct scans-a deep learning approach. In *International conference on medical image computing and computer-assisted intervention*, pages 678–686. Springer, 2015.

[96] A. Sekuboyina, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinitsch, J. S. Kirschke, and B. 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.

[97] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[98] D. Staub and M. J. Murphy. A digitally reconstructed radiograph algorithm calculated from first principles. *Medical physics*, 40(1):011902, 2013.

[99] A. Bayat, A. Sekuboyina, F. Hofmann, M. El Husseini, J. S. Kirschke, and B. H. Menze. Vertebral labelling in radiographs: Learning a coordinate corrector to enforce spinal shape. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, pages 39–46. Springer, 2019.

[100] J. Li, A. Pepe, C. Gsaxner, and J. Egger. An online platform for automatic skull defect restoration and cranial implant design. *arXiv:2006.00980*, 2020.

[101] M. Gall, X. Li, X. Chen, D. Schmalstieg, and J. Egger. Computer-aided planning and reconstruction of cranial 3d implants. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1179–1183, 08 2016.

[102] X. Chen, L. Xu, X. Li, and J. Egger. Computer-aided implant design for the restoration of cranial defects. *Scientific Reports*, 7:1–10, 06 2017.

[103] A. Marzola, L. Governi, L. Genitori, F. Mussa, Y. Volpe, and R. Furferi. A semi-automatic hybrid approach for defective skulls reconstruction. *Computer-Aided Design and Applications*, 17:190–204, 05 2019.

[104] J. Egger, M. Gall, A. Tax, M. Üçal, U. Zefferer, X. Li, G. von Campe, U. Schäfer, D. Schmalstieg, and X. Chen. Interactive reconstructions of cranial 3d implants under mevislab as an alternative to commercial planning software. *PLoS ONE*, 12:20, 03 2017.

[105] L. Angelo, P. Di Stefano, L. Governi, A. Marzola, and Y. Volpe. A robust and automatic method for the best symmetry plane detection of craniofacial skeletons. *Symmetry*, 11:245, 02 2019.

[106] A. Morais, J. Egger, and V. Alves. *Automated Computer-aided Design of Cranial Implants Using a Deep Volumetric Convolutional Denoising Autoencoder*, pages 151–160. 04 2019.

[107] M. Husseini, A. Sekuboyina, A. Bayat, B. H. Menze, M. Loeffler, and J. S. Kirschke. Conditioned variational auto-encoder for detecting osteoporotic vertebral fractures. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, pages 29–38. Springer, 2019.

[108] J. Li, A. Pepe, C. Gsaxner, G. von Campe, and J. Egger. A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. *arXiv preprint arXiv:2006.12449*, 2020.

[109] A. Dai, C. R. Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6545–6554, 2016.

[110] M. Sung, V. G. Kim, R. Angst, and L. J. Guibas. Data-driven structural priors for shape completion. *ACM Trans. Graph.*, 34:175:1–175:11, 2015.

[111] M. Sarmad, H. J. Lee, and Y. M. Kim. Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5891–5900, 2019.

[112] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 85–93, 2017.

[113] X. Hu, Y. Yan, W. Ren, H. Li, Y. Zhao, A. Bayat, and B. Menze. Feedback graph attention convolutional network for medical image enhancement. *arXiv preprint arXiv:2006.13863*, 2020.

[114] A. Bhowmik, S. Shit, and C. S. Seelamantula. Training-free, single-image super-resolution using a dynamic convolutional network. *IEEE Signal Processing Letters*, 25(1):85–89, 2017.

[115] I. Ezhov, T. Mot, S. Shit, J. Lipkova, J. C. Paetzold, F. Kofler, F. Navarro, M. Metz, B. Wiestler, and B. Menze. Real-time bayesian personalization via a learnable brain tumor growth model. *arXiv preprint arXiv:2009.04240*, 2020.

[116] F. Navarro, S. Shit, I. Ezhov, J. Paetzold, A. Gafita, J. C. Peeken, S. E. Combs, and B. H. Menze. Shape-aware complementary-task learning for multi-organ segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 620–627. Springer, 2019.

[117] U. Wickramasinghe, E. Remelli, G. Knott, and P. Fua. Voxel2mesh: 3d mesh model generation from volumetric data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 299–308. Springer, 2020.

[118] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–239. Springer, 2017.

[119] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[120] X. Yang, R. Kwitt, M. Styner, and M. Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[121] A. V. Dalca, E. Yu, P. Golland, B. Fischl, M. R. Sabuncu, and J. E. Iglesias. Unsupervised deep learning for bayesian brain mri segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 356–365. Springer, 2019.

[122] D. F. Pace, S. R. Aylward, and M. Niethammer. A locally adaptive regularization based on anisotropic diffusion for deformable image registration of sliding organs. *IEEE transactions on medical imaging*, 32(11):2114–2126, 2013.

[123] T. De Silva, A. Uneri, M. Ketcha, S. Reaungamornrat, G. Kleinszig, S. Vogt, N. Aygun, S. Lo, J. Wolinsky, and J. Siewerdsen. 3d–2d image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch. *Physics in Medicine & Biology*, 61(8):3009, 2016.

[124] A. Bayat, A. Sekuboyina, J. C. Paetzold, C. Payer, D. Stern, M. Urschler, J. S. Kirschke, and B. H. Menze. Inferring the 3d standing spine posture from 2d radiographs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 775–784. Springer, 2020.

[125] M. S. Nosrati and G. Hamarneh. Incorporating prior knowledge in medical image segmentation: a survey. *arXiv preprint arXiv:1607.01092*, 2016.

[126] D. Meng, M. Keller, E. Boyer, M. Black, and S. Pujades. Learning a statistical full spine model from partial observations. In *International Workshop on Shape in Medical Imaging*, pages 122–133. Springer, 2020.

[127] A. G. Roy, N. Navab, and C. Wachinger. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 421–429. Springer, 2018.

[128] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[129] A. Sekuboyina, M. Rempfler, A. Valentinitsch, B. H. Menze, and J. S. Kirschke. Labeling vertebrae with two-dimensional reformations of multidetector ct images: An adversarial approach for incorporating prior knowledge of spine anatomy. *Radiology: Artificial Intelligence*, 2(2):e190074, 2020.

[130] E. Konukoglu, B. Glocker, A. Criminisi, and K. M. Pohl. Wesd–weighted spectral distance for measuring shape dissimilarity. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2284–2297, 2012.

[131] E. Konukoglu, B. Glocker, D. H. Ye, A. Criminisi, and K. M. Pohl. Discriminative segmentation-based evaluation through shape dissimilarity. *IEEE transactions on medical imaging*, 31(12):2278–2289, 2012.