

Machine Learning with Structural Priors for Image Analysis of the Spine

Anjany Sekuboyina



Machine Learning with Structural Priors for Image Analysis of the Spine

Anjany Sekuboyina

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Daniel Cremers

Prüfende der Dissertation:

1. Prof. Dr. Bjoern H. Menze
2. Prof. Dr. Georg Langs
3. Prof. Dr. Martin Reuter

Die Dissertation wurde am 12.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 04.03.2024 angenommen.

అమ్మ కౌసం / To my mother



Abstract

Spine imaging, in the form of computed tomography (CT, among others), provides an insight into an essential structure of the human body. Automated extraction of information from spine images not only speeds up a radiologist’s workflow but also acts as a supporting-tool that identifies abnormalities in opportunistic scans (e.g. chest CT). The objective of this thesis is to automate this information-extraction from spine images in a data-driven manner.

We start by addressing the problem of localising and identifying the vertebrae, a fundamental step in any spine image processing pipeline. We propose an efficient convolutional neural network architecture that works on two-dimensional, orthogonal projections. This architecture achieves a detection performance comparable to three-dimensional architectures at a fraction of their computational budget. This network is further reinforced by an adversarial-learning regime that enforces an anatomical shape prior into the network’s prediction. We refer to this as *ad-hoc* prior learning. Next, we explore *post-hoc* enforcement of anatomical priors using a prior-informed linear conditional random field (CRF) that corrects the network’s predictions at inference-time. Solving a linear CRF during inference enables faster runtime, making its deployment feasible in a clinical setting. Additionally, we learn five different prior-models to account for anomalous spine anatomies, thereby achieving near 100% identification of vertebrae.

Following this, we shift our focus from automated processing to automated diagnosis in spine imaging. Here, we investigate vertebral fracture detection, also in a compute-parsimonious and annotation-limited setting. Assuming that an accurate vertebral segmentation mask is available, we task a generative model (variation auto-encoder) to learn the data distribution of healthy vertebrae in the point-cloud domain. A fractured vertebrae can then be identified as an outlier in this distribution. Furthermore, the model is extended to result in probabilistic reconstructions, which makes the detection of fractures interpretable by locating the region of the vertebra where the shape deviates from that of its healthy counterpart.

In conclusion, we have attempted to address some important questions towards automated spine image analysis, with a focus on the challenges in a clinical deployment such as compute restrictions, anatomical anomalies, and severe lack of

annotated data. Crucially, most of our answers have given rise to many other, more important questions in spine image processing. This thesis is an attempt to collate the questions, our answers, and the set of new questions that arose out of these answers.



Zusammenfassung

Die Wirbelsäulen-Bildgebung, unter anderem in Form von Computertomographie (CT), bietet einen Einblick in eine essentielle Struktur des menschlichen Körpers. Die automatisierte Extraktion von Informationen aus Wirbelsäulenbildern beschleunigt nicht nur den Arbeitsablauf eines Radiologen, sondern dient auch als Hilfsmittel zur Erkennung von Anomalien bei opportunistischen Scans (z. B. Brust-CT). Ziel dieser Arbeit ist es, diese Informationsextraktion aus Wirbelsäulenbildern auf daten-gesteuerte Weise zu automatisieren.

Wir beginnen mit dem Problem der Lokalisierung und Identifizierung der Wirbel, was als erster Schritt in jeder medizinischen Bildverarbeitungspipeline für die Wirbelsäule gilt. In diesem Zusammenhang wird eine effiziente Convolutional Neural Network-basierte Architektur, die orthogonale 2D-Projektionen verarbeitet, vorgestellt. Die vorgestellte Architektur demonstriert Erkennungsleistungen vergleichbar mit 3D Architekturen, während nur einen Bruchteil der Rechenressourcen benötigt wird.

Die Erkennungsleistung wird durch Training in einem adversen Lernregime (*adversarial learning-regime*) weiter verstärkt, indem eine anatomische Formpriorität in die Vorhersage eingebettet wird. Wir bezeichnen dies als *ad-hoc Prior Learning*. Zudem erforschen wir die *post-hoc* Durchsetzung von Prioritäten mit Hilfe eines priorinformierten *Conditional Random Fields* (CRF), das die Vorhersagen des Neuronalen Netzwerkes korrigiert. Das Lösen eines linearen CRFs während der Inferenz ermöglicht eine schnellere Laufzeit. Darüber hinaus lernen wir fünf verschiedene a-priori Modelle, um anomale Wirbelsäulenanatomien zu berücksichtigen, wodurch die Wirbelidentifikationsleistung auf nahezu 100% gesteigert wird.

Im Anschluss daran verlagern wir unseren Schwerpunkt von der automatisierten Verarbeitung auf die automatisierte Diagnose im Bereich der Wirbelsäulenbildgebung. Hier untersuchen wir die Erkennung von Wirbelfrakturen in einem ebenfalls rechenschwachen Umfeld unter Mangel an annotierten Daten. Unter der Annahme, dass eine genaue Wirbelsäulensegmentierung verfügbar ist, beauftragen wir ein generatives Modell (*Variational Autoencoder*) mit dem Erlernen der Datenverteilung gesunder Wirbel in der Punktwolkendomäne. Ein gebrochener Wirbel kann dann als ein Ausreißer in dieser Verteilung identifiziert werden. Darüber hinaus wird das

Modell so erweitert, dass es zu probabilistischen Rekonstruktionen führt, was die Erkennung von Frakturen durch die Lokalisierung der Wirbelregion, in der die Form von der eines gesunden Gegenstücks abweicht, interpretierbar macht.

Zusammenfassend haben wir versucht, wichtige Fragen zur automatisierten Wirbelsäulenbildanalyse zu beantworten, wobei wir uns auf die Herausforderungen im klinischen Einsatz konzentriert haben, wie z. B. Rechenbeschränkungen, anatomische Anomalien und einen erheblichen Mangel an annotierten Daten. Entscheidend ist, dass die meisten unserer Antworten zu vielen anderen, noch wichtigeren Fragen in der Wirbelsäulenbildverarbeitung geführt haben. Diese Arbeit ist ein Versuch, die Fragen, unsere Antworten und eine Reihe neuer Fragen, die sich aus diesen Antworten ergeben haben, zusammenzustellen.



Acknowledgements

Foremost, I acknowledge my mother. All she had to say was ‘No’ and my doctoral journey would not even have started. She knew this, and she had all the reasons (and more) to say *no*. But, she said *yes* and set me off on this special journey. Thank you, Amma, for everything.

No less important in this journey of mine are my two supervisors, Jan Kirschke and Bjoern Menze. Jan, thank you for the immense support you’ve provided me and for all the trust you’ve put in me, all these years. Thank you for letting me explore and for guiding me back if I strayed too far. Thank you for holding my hand and leading me to build something I would’ve never imagined to. Bjoern, thank you for all the freedoms you have given me in problem solving, especially for teaching me how to tread that fine line between applied and clinical research. Thank you for being patient with my research failures. Importantly, thank you for always encouraging me to strive for clinical translation for my research. I couldn’t have asked for a better pair of supervisors.

My doctoral journey has been filled with incredible people, both at work and outside it. I thank Markus, Alex, Jan, Dhritiman, Cagdas, Jana, and Esther for helping me navigate a new workplace, a new country, and a new culture. I explicitly thank Fernando, Suprosanna, Johannes, Ivan, Giles, and Bran, for showing me a new perspective towards not only work but also life. Effective towards whoever I’m today are the roles played by many individuals on a similar journey, Malek, Chinmay, Tamaz, Amir, Carolin, Oliver, Florian, Judith, Diana, and Marie. Thank you. And here is to Jan, Giles, Malek, Sebastian, and Dominik for accompanying me in this extremely rewarding and highly uncertain journey. Thank you, Sravan, Talia, and Sirisha, for making this journey so full of beautiful memories. When I think back to the last few years, these memories are the one of the first that my mind brings up.

I’d like to gratefully acknowledge Prof. Dr. Daniel Cremers for chairing my dissertation and to Prof. Dr. Georg Langs for examining it. I also thank Brandon and Daniel for offering me a glimpse into research in the industry and for making my stay in Heidelberg enjoyable.

Lastly, I am forever indebted to my father, my grandfather, and my big and

beautiful family back home. They have always loved me unconditionally, encouraged me in all my endeavours, and have been proud of whatever little I've achieved. Please know that I'm more proud of you all, and I have always thanked this universe for putting me in your fold.



Contents

Abstract	iii
Zusammenfassung	v
Acknowledgements	vii
Contents	ix
List of Figures	xiii
Publication List	xv
I INTRODUCTION	1
1 Foreword	3
2 Background	7
2.1 Anatomy of the spine	7
2.2 Spine analysis in a clinical setting	8
2.3 Imaging modalities for the spine	10
2.3.1 Computed Tomography	10
2.3.2 Magnetic Resonance Imaging	12
3 Methodology	13
3.1 Preliminaries: Data, Models, and Losses	14
3.1.1 Model Architecture	15
3.1.2 Loss	16
3.2 Generative models	18
3.2.1 Generative Adversarial Networks	18
3.2.2 Variational Autoencoders	20

CONTENTS

3.3	Landmark detection and anatomical priors	22
3.4	Point clouds and anomaly detection	24
4	Summary of the Contributions	27
II PUBLICATIONS		29
5	Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior	31
6	Labeling Vertebrae with Two-dimensional Reformations of Multi-detector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy	41
7	Pushing the limits of an FCN and a CRF towards near-ideal vertebrae labelling	53
8	Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis	61
III CONCLUDING REMARKS		71
9	Discussion	73
10	Outlook	77
	Bibliography	81
IV APPENDICES		89
A	Supplementary Material: Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior	91

B Supplementary Material: Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy . . . **95**

C Supplementary Material: Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis **101**

D VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images **105**

E Anduin: An open-source, web-based spine segmentation tool . . . **139**

F Anduin: An open-source, web-based spine segmentation tool . . . **141**



List of Figures

- 2.1 (a) Anatomy of a typical human spine. (b) Sub-regions of the various vertebrae in the spine 8
- 2.2 Diversity of spine CTs in a clinical setting(Labelled clockwise) Algorithms analysis the spine must be robust to wide variation of spine scans in terms of: the fields of view, fractured vertebrae (B, J), metal insertions (C), cemented vertebrae (G), transitional vertebrae (L6 and T13 in D and I respectively), scan noise (K) etc. Source: [4] 9
- 3.1 A pictorial description of a spine pipeline for spine image analysis 13
- 3.2 Architectural illustration of three fundamental *deep* neural networks. 15
- 3.3 Block diagram of a Generative adversarial network. 18
- 3.4 Block diagram of a Wasserstein GAN. 19
- 3.5 Block diagram of an Energy-based GAN. 20
- 3.6 Block diagram of an autoencoder and a variational autoencoder. 21
- 3.7 **Landmark detection:** Two traditional approaches towards landmark detection: (a) coordinate-regression (b) heatmap-regression 22
- 3.8 **Priors in landmark detection:** Two traditional approaches towards landmark detection: (a) coordinate-regression (b) heatmap-regression 23
- 3.9 An illustration describing outlier-detection in the point-cloud domain 24
- F.1 **Anduin’s landing page:** Anduin is hosted at `anduin.bonescreen.de`, released under CC-BY-SA 4.0 license. Users have the ability to request for user accounts after agreeing to the data policy along with anduin’s terms and conditions on data processing 142
- F.2 **Upload dialogue:** Users have the ability to upload NIFTI files (*.nii or *.nii.gz) along with a JSON side car containing the header information 143
- F.3 **Processing status:** Once uploaded, the scan can be processed to result in an ‘evaluation’ containing the processing artefacts. Processing of a CT of diagnostic quality takes approximately 1–2 minutes. 143

LIST OF FIGURES

- F.4 **Artefact download:** Once processed, every artefact of the processing pipeline can be download, as required. This includes the bounding box around the spine, the vertebral centroids, the segmentation masks or the vertebrae and their subregion masks. 144
- F.5 A snapshot of *anduin*'s output, the left two tiles showing the sagittal and coronal reformations with the predicted centroids and segmentation masks and the two tiles on the right showing the maximum intensity projection of the subregion mask (sagittal and coronal). 144



Publication List

The following four publications constitute the core of my *cumulative doctoral thesis*. A * indicates shared first authorship.

- [1] **A. Sekuboyina**, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinitich, J. S. Kirschke, and B. H. Menze. “Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior.” In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 649–657.
- [2] **A. Sekuboyina**, M. Rempfler, A. Valentinitich, 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.” In: *Radiology: Artificial Intelligence 2.2* (2020), e190074.
- [3] **A. Sekuboyina**, M. Rempfler, A. Valentinitich, M. Loeffler, J. S. Kirschke, and B. H. Menze. “Probabilistic point cloud reconstructions for vertebral shape analysis.” In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer. 2019, pp. 375–383.
- [4] **A. Sekuboyina***, J. Irmay*, S. Shit, J. Kirschke, B. Andres, and B. Menze. “Pushing the limits of an FCN and a CRF towards near-ideal vertebrae labelling.” In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5.

The following additional 29 publications were further co-authored *during the time of my doctoral thesis* with significant contribution. This includes contribution to both methodological and clinical aspects, as a direct consequence of work done in this thesis. A * indicates shared first authorship.

2022

- [1] C. Prabhakar, **A. Sekuboyina**, S. Shit, et al. “HeteroKG: Knowledge graphs for multi-modal learning.” In: *Geometric Deep Learning in Medical Image Analysis (Extended abstracts)*. 2022.
- [2] C. Prabhakar, **A. Sekuboyina**, H. B. Li, J. C. Paetzold, S. Shit, T. Amiranashvili, J. Kleesiek, and B. Menze. “Structured Knowledge Graphs for Classifying Unseen Patterns in Radiographs.” In: *Geometric Deep Learning in Medical Image Analysis*. PMLR. 2022, pp. 45–60.
- [3] S. Rühling, F. Navarro, **A. Sekuboyina**, M. El Hussein, T. Baum, B. Menze, R. Braren, C. Zimmer, and J. S. Kirschke. “Automated detection of the contrast phase in MDCT by an artificial neural network improves the accuracy of opportunistic bone mineral density measurements.” In: *European Radiology* (2022), pp. 1–10.
- [4] A. Bayat, D. F. Pace, **A. Sekuboyina**, C. Payer, D. Stern, M. Urschler, J. S. Kirschke, and B. H. Menze. “Anatomy-aware inference of the 3D standing spine posture from 2D radiographs.” In: *Tomography* 8.1 (2022), pp. 479–496.
- [5] S. Rühling, A. Scharr, N. Sollmann, M. Wostrack, M. T. Löffler, B. Menze, **A. Sekuboyina**, M. El Hussein, R. Braren, C. Zimmer, et al. “Proposed diagnostic volumetric bone mineral density thresholds for osteoporosis and osteopenia at the cervicothoracic spine in correlation to the lumbar spine.” In: *European Radiology* 32.9 (2022), pp. 6207–6214.
- [6] N. Sollmann, M. T. Löffler, M. El Hussein, **A. Sekuboyina**, M. Dieckmeyer, S. Rühling, C. Zimmer, B. Menze, G. B. Joseph, T. Baum, et al. “Automated opportunistic osteoporosis screening in routine computed tomography of the spine: comparison with dedicated quantitative CT.” In: *Journal of Bone and Mineral Research* 37.7 (2022), pp. 1287–1296.

-
- [7] M. Dieckmeyer, N. Sollmann, M. El Hussein, **A. Sekuboyina**, M. T. Löffler, C. Zimmer, J. S. Kirschke, K. Subburaj, and T. Baum. “Gender-, age- and region-specific characterization of vertebral bone microstructure through automated segmentation and 3D texture analysis of routine abdominal CT.” In: *Frontiers in Endocrinology* 12 (2022), p. 1956.
- [8] T. Lerchl, M. El Hussein, A. Bayat, **A. Sekuboyina**, L. Hermann, K. Nispel, T. Baum, M. T. Löffler, V. Senner, and J. S. Kirschke. “Validation of a Patient-Specific Musculoskeletal Model for Lumbar Load Estimation Generated by an Automated Pipeline From Whole Body CT.” In: *Frontiers in Bioengineering and Biotechnology* 10 (2022).
- [9] M. Dieckmeyer, M. T. Löffler, M. El Hussein, **A. Sekuboyina**, B. Menze, N. Sollmann, M. Wostrack, C. Zimmer, T. Baum, and J. S. Kirschke. “Level-specific volumetric BMD threshold values for the prediction of incident vertebral fractures using opportunistic QCT: a case-control study.” In: *Frontiers in Endocrinology* 13 (2022).

2021

- [1] **A. Sekuboyina**, M. E. Hussein, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern, et al. “VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images.” In: *Medical image analysis* 73 (2021), p. 102166.
- [2] **A. Sekuboyina**, D. Oñoro-Rubio, J. Kleesiek, and B. Malone. “A relational-learning perspective to multi-label chest X-ray classification.” In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1618–1622.
- [3] H. Liebl, D. Schinz, **A. Sekuboyina**, L. Malagutti, M. T. Löffler, A. Bayat, M. El Hussein, G. Tetteh, K. Grau, E. Niederreiter, et al. “A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data.” In: *Scientific Data* 8.1 (2021), p. 284.

- [4] S. Shit*, J. C. Paetzold*, **A. Sekuboyina**, I. Ezhov, A. Unger, A. Zhylyka, J. P. Plum, U. Bauer, and B. H. Menze. “cIDice-a Novel Topology-Preserving Loss Function for Tubular Structure Segmentation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16560–16569.
- [5] H. Li, R. G. Prasad, **A. Sekuboyina**, C. Niu, S. Bai, W. Hemmert, and B. Menze. “Micro-Ct Synthesis and Inner Ear Super Resolution via Generative Adversarial Networks and Bayesian Inference.” In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1500–1504.
- [6] H. Li, S. Gopal, **A. Sekuboyina**, J. Zhang, C. Niu, C. Pirkl, J. Kirschke, B. Wiestler, and B. Menze. “Unpaired MR image homogenisation by disentangled representations and its uncertainty.” In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. Springer. 2021, pp. 44–53.
- [7] M. Dieckmeyer, N. M. Rayudu, L. Y. Yeung, M. Löffler, **A. Sekuboyina**, E. Burian, N. Sollmann, J. S. Kirschke, T. Baum, and K. Subburaj. “Prediction of incident vertebral fractures in routine MDCT: Comparison of global texture features, 3D finite element parameters and volumetric BMD.” In: *European Journal of Radiology* 141 (2021), p. 109827.
- [8] M. T. Löffler, A. Jacob, A. Scharr, N. Sollmann, E. Burian, M. El Hussein, **A. Sekuboyina**, G. Tetteh, C. Zimmer, J. Gempt, et al. “Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA.” In: *European radiology* 31 (2021), pp. 6069–6077.
- [9] N. Sollmann, N. M. Rayudu, L. Y. Yeung, **A. Sekuboyina**, E. Burian, M. Dieckmeyer, M. T. Löffler, B. J. Schwaiger, A. S. Gersing, J. S. Kirschke, et al. “MDCT-based finite element analyses: are measurements at the lumbar spine associated with the biomechanical strength of functional spinal units of incidental osteoporotic fractures along the thoracolumbar spine?” In: *Diagnostics* 11.3 (2021), p. 455.
- [10] L. Y. Yeung, N. M. Rayudu, M. Löffler, **A. Sekuboyina**, E. Burian, N. Sollmann, M. Dieckmeyer, T. Greve, J. S. Kirschke, K. Subburaj, et al. “Prediction of incidental osteoporotic fractures at vertebral-specific level using 3D

non-linear finite element parameters derived from routine abdominal MDCT.”
In: *Diagnostics* 11.2 (2021), p. 208.

2020

- [1] M. T. Löffler, **A. Sekuboyina**, A. Jacob, A.-L. Grau, A. Scharr, M. El Hussein, M. Kallweit, C. Zimmer, T. Baum, and J. S. Kirschke. “A vertebral segmentation dataset with fracture grading.” In: *Radiology: Artificial Intelligence* 2.4 (2020), e190138.
- [2] M. Hussein, **A. Sekuboyina**, A. Bayat, B. H. Menze, M. Loeffler, and J. S. Kirschke. “Conditioned variational auto-encoder for detecting osteoporotic vertebral fractures.” In: *Computational Methods and Clinical Applications for Spine Imaging: 6th International Workshop and Challenge, CSI 2019, Shenzhen, China, October 17, 2019, Proceedings 6*. Springer. 2020, pp. 29–38.
- [3] M. Hussein*, **A. Sekuboyina***, M. Loeffler, F. Navarro, B. H. Menze, and J. S. Kirschke. “Grading loss: a fracture grade-based metric loss for vertebral fracture detection.” In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer. 2020, pp. 733–742.
- [4] 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: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer. 2020, pp. 775–784.
- [5] A. Bayat, **A. Sekuboyina**, F. Hofmann, M. E. Hussein, J. S. Kirschke, and B. H. Menze. “Vertebral labelling in radiographs: learning a coordinate corrector to enforce spinal shape.” In: *Computational Methods and Clinical Applications for Spine Imaging: 6th International Workshop and Challenge, CSI 2019, Shenzhen, China, October 17, 2019, Proceedings 6*. Springer. 2020, pp. 39–46.
- [6] F. Navarro, **A. Sekuboyina**, D. Waldmannstetter, J. C. Peeken, S. E. Combs, and B. H. Menze. “Deep reinforcement learning for organ localization in CT.” In: *Medical Imaging with Deep Learning*. PMLR. 2020, pp. 544–554.

2019

- [1] H. Li, J. C. Paetzold, **A. Sekuboyina**, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze. “DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis.” In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 795–803.
- [2] Y. Zhao, H. Li, S. Wan, **A. Sekuboyina**, X. Hu, G. Tetteh, M. Piraud, and B. Menze. “Knowledge-aided convolutional neural network for small organ segmentation.” In: *IEEE journal of biomedical and health informatics* 23.4 (2019), pp. 1363–1373.

2018

- [1] M. Piraud, **A. Sekuboyina**, and B. H. Menze. “Multi-level activation for segmentation of hierarchically-nested classes.” In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018, pp. 0–0.
- [2] **A. Sekuboyina***, J. Kukačka*, J. S. Kirschke, B. H. Menze, and A. Valentinitsch. “Attention-driven deep learning for pathological spine segmentation.” In: *Computational Methods and Clinical Applications in Musculoskeletal Imaging: 5th International Workshop, MSKI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Revised Selected Papers 5*. Springer. 2018, pp. 108–119.

PART I

INTRODUCTION



Foreword

Look well to the **spine** for the cause of disease.

Hippocrates
Father of western medicine

The human spine is responsible for the human posture and support, holding almost every organ in place. It protects the spinal cord, the neural highway of the body. The vertebrae, which are the individual bones of the spine, protect the spinal cord. Along with the inter-vertebral discs, the vertebrae bear majority of the weight on the spine. The neural foramen are where the nerve roots exit the spinal canal to the rest of the body. The paraspinal muscles are responsible crucial motion of the body, and so on. Thus, the spine is one of the most critical and complex anatomies of the human body. However, all is not well in the spine. Back pain is one of the most commonly reported problems in the world. Von der Lippe et al. [1] report that more than 60% of the respondents in a study in Germany reported back pain in 2020. In UK, its prevalence is around 80% [2]. Osteoporosis, a disease characterised by low bone mass leading to frailty, is extremely common worldwide. More than 32 million in Europe and more than 10 million people in the USA are effected by osteoporosis today. Annually, around 4.3 million osteoporotic fractures will occur, resulting in a debilitating effect in the quality of life. It also leads to an 8-fold higher mortality rate [3]. Most common malignant cancers predominantly metastasise to the spine, e.g. breast cancer (21%), lung cancer (19%), etc. Therefore, it is of immense interest to look at the spine.

Every year, around 66 million CT scans are imaged in Europe. Nearly 50% of the capture some part of the spine. However, less than 1% of all the CT scans are actually imaged to screen the spine. This is wasted opportunity cost. Now consider the fact that a chest/abdomen/pelvis CT takes the longest to *read* compared to say a brain CT or a chest CT. In a setting where the radiologist is already constrained for time, the long reading times further exacerbate the cost of this wasted opportunity.

Eventually, for instance, close to 90% of osteoporotic fractures go undiagnosed or significant fraction of bone cancers are misdiagnosed. Hence, there is a clear need to design computer-aided, support systems for radiological reading of spine images.

One fundamental task of any computer-aided support system is to *understand* the raw image data, which is broadly referred to medical image analysis. In other words, this involves *abstracting* information from the raw image data, thereby extracting useful information for subsequent reading. A primary task in medical image analysis is **anatomical landmark detection**. In the context of the spine, this refers to the challenging task *vertebrae labelling* which involves localising and identifying the vertebrae. Labelling the vertebrae have immediate diagnostic consequences such as estimating the spinal curvature and identifying deformities such as scoliosis and kyphosis. From a non-diagnostic perspective, it enables important downstream tasks such as vertebral segmentation, fracture detection, surgical planning, and biomechanical modelling. One task immediately enabled by vertebral labelling is vertebral segmentation [4, 5], which in turn plays a crucial role in **vertebral fracture detection** and grading [6]. As stated above, vertebral fractures are extremely common and almost certainly crippling.

Deep neural networks (DNN) are ubiquitous not only in natural image processing but also on medical image processing. They have been used for numerous tasks such as disease classification, anatomical segmentation, and anatomical landmark detection. So, what makes spine image processing special? Typical spine scans are very large ($\sim 10^6$ voxels at a spatial resolution of 1mm^3), which conventional DNNs struggle to process readily. Moreover, the scans come in all shapes and forms in terms of fields-of-view, metal insertions, anatomical anomalies such as transitional vertebrae, fractured vertebrae etc. However, the number of naturally occurring cases with these exceptions are few. This severe data imbalance calls for novel approaches towards handling the tasks of spine image processing. Finally, if our objective is to design a system that can be deployed in a real clinical setting, one needs to also consider the run-time, compute budget, the system’s explainability, etc.

A typical spine image analysis system consists of two parts: an *image-to-structure* component which is responsible for giving raw image some structure in the form of landmarks, segmentation masks etc., and a *structure-to-diagnosis component* which includes any downstream task that the structure could enable, e.g. detect fractures from vertebral segmentation masks. This thesis deals with both these components of the system with a focus on designing algorithms are based on learning *anatomical data priors*. First, the thesis presents methods aimed at labelling vertebrae by enforcing local and global anatomical priors into the feed-forward network using ad-

versarial learning or conditional random-fields (CRF). The presented methods are designed to be accurate, robust, and deployable on low-resource medical hardware. Next, given a structured spine image (e.g. labelled and segmented), the thesis tackles the downstream task of fracture detection by posing it as an *out-of-distribution detection* (OOD) problem. Such formulation works around the issue of data-imbalance described above, as the system can now be trained only on healthy vertebrae. The presented method, working in the point-cloud domain, results in fast inference times and requires little compute resources. As mentioned above, translation of machine learning into real-world clinical setting is a recurring theme of this thesis.

Organisation

This dissertation consists of three parts, each part divided into chapters. Part I, of which the this Foreword (1) chapter is a part, consists of three other chapters. Chapter 2 lays out the medical background required to understand and appreciate this dissertation, mostly dealing with the anatomy of the spine, the challenges, and the current trends and so on. Following this, a review of the the key methodological concepts employed in this dissertation is given in 3. As a final part of introduction, Chapter 4 collects the open questions described prior and summarises the contributions of this thesis. The second part of this thesis (Part II) consists of four peer-reviewed publications that constitute this thesis, each as a chapter, Chapters 5-8. Every publication is self-contained in terms of introduction, methodology, experiments, and discussion. However, each publication is preceded by a brief synopsis on how the publication fits into the bigger scheme of this dissertation. Lastly, in Part III, we discuss the presented work in Chapter 9 and draw an overall conclusion along with an outlook for the future in Chapter 10. Finally, Part IV consists of the appendix with two related, high-impact works that are a consequence of this thesis: (1) VERSE, a large-vertebrae segmentation challenge, and (2) <https://andu.in.bonescreen.de>, an open-source web application for spine segmentation, which speak towards the clinical translation often touched upon in this thesis.



Background

This chapter provides a brief overview of the anatomy of the spine and the imaging modality (CT) explored in this thesis. Thereby, it also describes the challenges that are faced during real-world deployment of any spine-based image-processing algorithms, with the objective of helping the reader appreciate the complexity involved in automated spine image analysis.

2.1 Anatomy of the spine

A typical human spine is an *S*-shaped stack of 33 individual bones. As shown in Fig. 2.1a, spine is divided into three major regions: the cervical spine, the thoracic spine, and the lumbar spine. Typically, each of these regions consists of seven (C1–C7), twelve (T1–T12), and five (L1–L5) vertebrae respectively. Additionally, the spine also consists of five sacral vertebrae (S1–S5) forming the sacrum and one coccyx. The first 24 vertebrae are movable while the sacrum and the coccyx are fused to the hip bone and immovable. The first part of this thesis deals with automatically localising these vertebrae in a scan and identifying them. Between a pair of every movable vertebra is the inter-vertebral disc (IVD), a gel-filled disc which cushions the vertebrae and prevents their rubbing. Note that if the vertebrae are identified, the identification of the IVDs naturally follows.

Fig. 2.1b shows the large variability in the shapes of the individual vertebrae. Almost every vertebra can be broadly sub-divided into three parts: (1) the vertebral body or the anterior, (2) the vertebral arch forming the spinal canal, and (3) posterior, consisting of the spinous and transverse processes, and facets etc. Observe that C1 and C2 are exceptions to this and have a special morphology. The second part of this thesis aims to characterise the morphology of these vertebrae in order to identify outlying shapes.

There exist other, very critical, components of the spine which are not looked into as part of this thesis. The spinal muscles, majorly the extensors and the flexors, aid human motion. The former are also referred to as the back muscles, crucial for

2. BACKGROUND

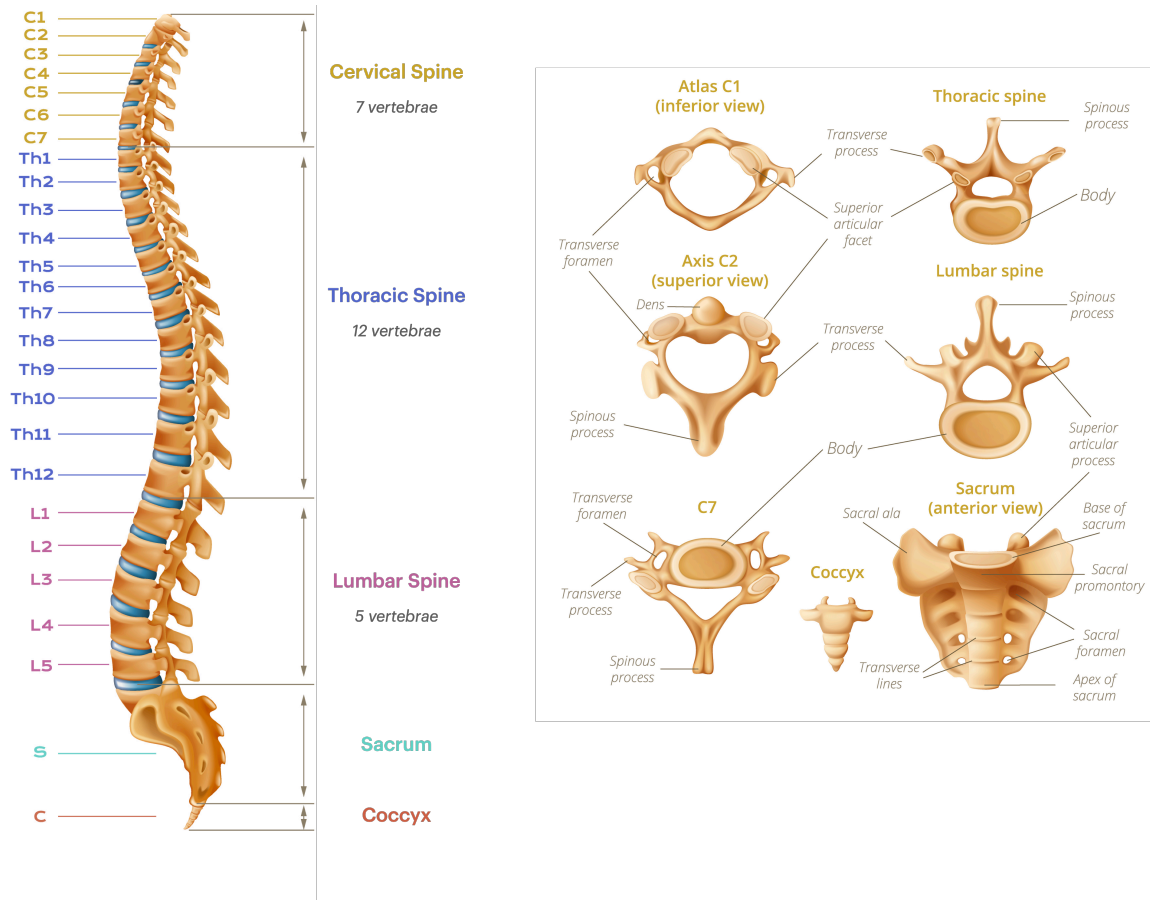


Figure 2.1: (a) Anatomy of a typical human spine. (b) Sub-regions of the various vertebrae in the spine

spine stabilisation, and the site where back pain is reported. The spinal cord that runs down the entire spinal canal is the information super-highway, relaying signals from the brain to the rest of the body, branching into spinal nerves at the vertebral foramen.

2.2 Spine analysis in a clinical setting

When deploying algorithms for automated spine analysis, it is essential that they not only work for typical, healthy spines but also for abnormal spines. In this subsection,

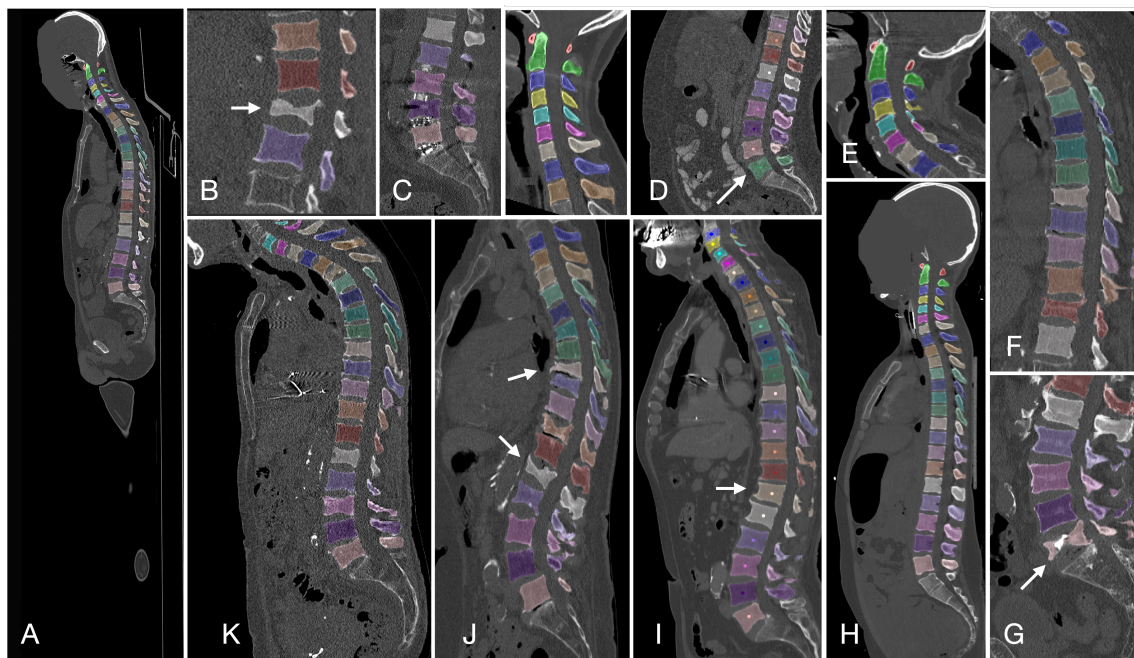


Figure 2.2: Diversity of spine CTs in a clinical setting (Labelled clockwise) Algorithms analysis the spine must be robust to wide variation of spine scans in terms of: the fields of view, fractured vertebrae (B, J), metal insertions (C), cemented vertebrae (G), transitional vertebrae (L6 and T13 in D and I respectively), scan noise (K) etc. Source: [4]

we list of set of challenging cases that the an algorithm is bound to face (due to their natural prevalence) when working with the spine. Fig. 2.2 shows a (non-exhaustive) overview of the scans an algorithm would have to process.

- **Field-of-view.** Spine is one of the largest component in the body. As a consequence, it is imaged in a variety of scans, e.g a cardiac MR, a lung CT, abdomen scan etc. This variety in the FoV of the scans must be efficiently tackled.
- **Spines with degeneration.** With patient's age, the prevalence of degeneration in the spine increases [7]. Degeneration could occur at the level of the spine (e.g. kyphosis or lordosis) or at the level of the vertebra (e.g. fractures). Additionally, these degenerations might also have been *addressed* by inserting metal plates, screws, or cement. The algorithms should thus be robust to degeneration as well

as its treatment.

- **Anatomical abnormalities.** In Sec. 2.1, we learnt that that a typical spine consists of twelve thoracic vertebrae and five lumbar vertebrae. However, a significant fraction of scans seen in clinical settings contain an atypical anatomy. Following are a few anatomical abnormalities that could exist in a clinical setting:
 1. An additional thoracic vertebrae, T13.
 2. Missing T12.
 3. An additional lumbar vertebra, L6. Lumbosacral transitional vertebra where S1 assimilates to a lumbar vertebra.
 4. Missing L5, possibly caused by its sacralisation.

It should be noted that the prevalence of these anomalies is not insignificant (e.g. 35% of scans could have a transitional lumbosacral vertebrae) and should be accounted for when deploying algorithms into the clinic.

- **Efficient computation.** Medical images are three-dimensional and the extent of spine is large. When working at an isotropic spatial resolution of 1mm, this could result in a scan sizes between 10^6 to 10^8 voxels (two orders higher than a natural image of size 1024×1024). Hence, computational efficiency of to be taken into account for a clinical deployment. This could involve opting for light-weight data modalities, pipelining methods to remove rudimentary image information, or model simplification (e.g. linear vs. quadratic).

2.3 Imaging modalities for the spine

Having an overview of the spinal anatomy, we now look into how this anatomy can be imaged, especially 2D radiography and 3D magnetic resonance imaging (MRI), and 3D computed tomography (CT)

2.3.1 Computed Tomography

X-rays are a form of electromagnetic radiation that have an energy higher than visible light and a wavelength much smaller than it. As a result, they can penetrate and pass

through many objects, including the human body. However, the x-rays can also be absorbed by the human tissue at a rate dependant on the tissue type. The resulting, un-absorbed rays, on hitting the detector places on the other side of the body, form the image. A 2D x-ray radiograph, therefore, is an image of the *shadow* of the internal tissue structure. Computed tomography (CT) imaging elevates 2D radiography to 3D by employing a rotating x-ray tube (source) and multiple x-ray detectors in a spiral gantry that measure the x-ray attenuation or absorption. This results in multiple ‘virtual’ x-ray projections of the human body from various views. The reconstruction of a 3D CT image from these multiple but finite number of projections is a multi-dimensional inverse problem also called *tomographic reconstruction*. Since x-rays travel easily through less dense regions such as muscle and air-cavities and face high attenuation in dense tissue such as bones, radiographic imaging is useful for imaging bones, calcification, foreign objects, etc. Since the theme of this thesis is to localise the vertebrae and and to analyse its morphology, CT is chosen as the modality-of-interest.

Hounsfield Units. As stated, the voxel corresponding to an anatomical region in a CT image correspond to the mean x-ray attenuation of the tissue in that region, in other words, its radiodensity. This is represented quantitatively using ‘Hounsfield Unit (HU) Scale’, between -1024 (least attenuating) to + 3071 (most attenuating). HU is defined relative to the attenuation coefficients of air (-1024 HU) and water (0 HU). Attenuation of bone typically lies between +400 HU to +2000 HU. The process of increasing the dynamic range of the CT to better emphasise the densities within a certain range ‘windowing’. E.g. A bone window between 500 HU to 2000 HU will emphasise the bone. In this thesis, all data is calibrated to the Housfield Scale, which enables efficient data re-normalisation.

Multi-planar reconstructions. Recall that there is a need to explore efficient data representations for efficient processing of information in spine scans. An example of such representations are Multi-planar reconstructions (MPR). For instance, the voxel in CT represents the *mean* attenuation of an anatomical region. This is referred to as the average intensity projection (AIP). However, the same voxel can also represent the *maximum* attenuation in an anatomical region. This would result in the maximum intensity projection (MIP). The minimum intensity projection (minIP) can be defined by extension. Different projections highlight regions with specific attenuation, e.g. AIP is used for investigating solid organs, MIP for spine [8] or angiographic studies, and MinIP for air spaces such as lungs.

2.3.2 Magnetic Resonance Imaging

For the sake of completion, we describe another common imaging modality for the spine, Magnetic Resonance Imaging (MRI). Unlike x-rays, MR imaging does not work with radiation. It works by re-aligning the spins of the protons in the hydrogen atoms using magnetic fields. The abundance of water in human body makes MR imaging possible. Radio-frequency pulses are used to deflect the spins of the photons, which release energy as they return to their resting alignment. This energy is characteristic of the tissue type and is captured by the receiver coils, called k -space measurements. An inverse Fourier transform of these measurements results in the MR image.

Due to its dependency on a tissue's water content, MRI is usually unsuitable for bone imaging. However, it is predominantly used for imaging soft-tissue, which in the realm of spine, includes the sub-structures such as the spinal cord, the IVDs, the spinal muscles, tumours etc. Since spine extends over a large region, spine MR imaging is time-intensive. Typically spine MR images are limited either in terms of anatomical FoV (e.g. cervical spine or lumbar spine) or in terms of resolution (e.g. sagittal reformation or axial reformation) depending on the purpose of the scan. In Chapter 9, we allude to this limitation for suggesting future work in the domain of MR image analysis.



Methodology

Any medical image processing pipeline can be split into two-stages: First, extract a meaningful representation or structure from the image and second, use the representation to perform any downstream task such as diagnosis. A typical spine processing pipeline is illustrated in Fig. 3.2. Of interest for this thesis are the subjects of vertebrae labelling and fracture detection, both of which are approached from a common perspective: anatomical prior learning and efficient data representations for faster computation.

In this section, we provide the methodological background essential towards understanding the contributions of this thesis towards the two topics mentioned above. We start by motivating fundamental deep learning models and eventually build towards generative models which are employed for prior-learning. We also introduce the representation of data in the point cloud domain and motivate outlier detection

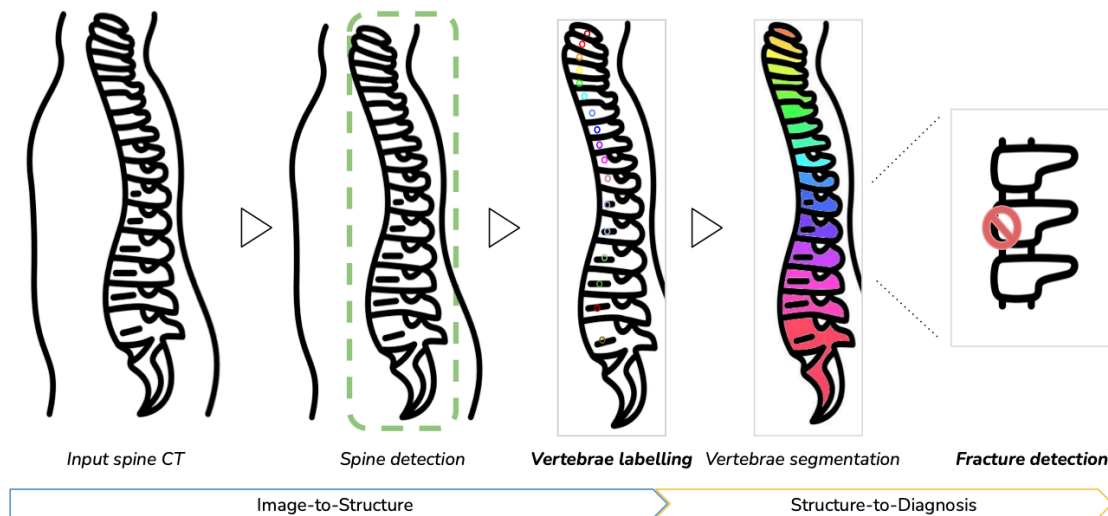


Figure 3.1: A pictorial description of a spine pipeline for spine image analysis

in this domain, which is the theme of this thesis’ contribution towards fracture detection. Please note that the description is intended to provide a concise overview and hence takes the liberty of simplifying the concepts.

3.1 Preliminaries: Data, Models, and Losses

Tom M. Mitchell (1997) defines a learning system as:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

In a data-driven approach, the experience E is typically gained using an annotated dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$. The input $x \in \mathcal{X}$ and its annotation $y \in \mathcal{Y}$ vary with task T . x could be a feature vector (\mathbb{R}^d), an image ($\mathbb{R}^{H \times W}$), a 3D CT scan ($\mathbb{R}^{H \times W \times D}$), or a point-cloud in 3D space ($\mathbb{R}^{N \times 3}$, where N is the number of surface points). Similarly, the dimensionality of y is defined by the task T .

Given \mathcal{D} , the aim of learning is to learnt a mapping \mathcal{F} parameterised by θ such that:

$$\mathcal{F}_\theta : \mathcal{X} \rightarrow \mathcal{Y} \tag{3.1}$$

The function is learnt by minimising a loss criterion \mathcal{L} over the space of parameter space such that:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\mathcal{F}_\theta(x), y), \forall \{x, y\} \in \mathcal{D}, \tag{3.2}$$

Of interest in this thesis are neural networks, which offer one way to model \mathcal{F}_θ . The field of *deep learning*, led by AlexNet [9], is currently ubiquitous and has achieved super-human performance in many vision-based tasks. In the following section, we briefly provide of an overview of some deep neural network (DNN) architectures. \mathcal{F}_θ is a DNN if it can be decomposed into a composition of functions:

$$\mathcal{F}_\theta(x) = (\mathcal{f}_\theta^1 \cdot \mathcal{f}_\theta^2 \cdot \dots \cdot \mathcal{f}_\theta^n)(x), \tag{3.3}$$

where \mathcal{f}_θ^i denotes a non-linear function and the choice of modelling it defines the DNN architectures.

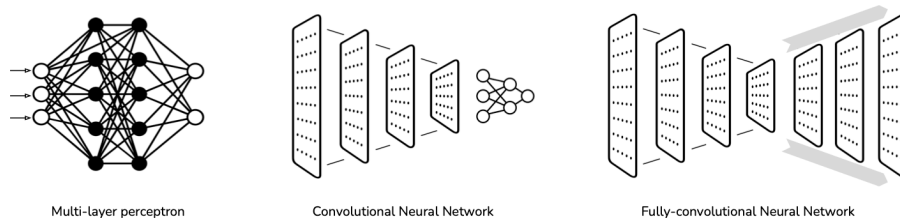


Figure 3.2: Architectural illustration of three fundamental *deep* neural networks.

3.1.1 Model Architecture

Broadly, DNNs are a composition of encoders and decoders. The role of an encoder is to take the high-dimensional data as input and *encode* them into a compact *representation*, while the role of a decoder is to generate high-dimensional data from the encoded representation. In this work, we focus on three relevant network architectures: multi-layer perceptrons, convolutional neural networks, and the fully-convolutional neural networks.

Multi-layer perceptron

A Multi-Layered Perceptron (MLP) introduced by Rosenblatt et al. [10] is a densely-connected neural network for when the data is vector-valued, i.e. $x \in \mathbb{R}^H$. A layer of an MLP can be represented as $f_{\theta}^i(x) = W_i x + b_i$, where $W_1 \in \mathbb{R}^{H \times h}$ is the linear transformation matrix and b_i is the bias. If $h < H$, the resulting representation is an encoded version of the input. A two-layered perceptron with a non-linearity (σ) can be described as:

$$\mathcal{F}_{MLP}(x) = W_2(\sigma(W_1 x + b_1)) + b_2 \quad (3.4)$$

Convolutional neural networks

Aimed at efficiently extracting local patterns in images by exploiting spatial-invariance, convolutional neural networks (CNN) were first introduced in [11]. A layer in a CNN is a filter that is convolved with the input (instead of densely multiplied as in an MLP). Moreover, *encoding* or reduction in feature dimension in a CNN is achieved using downsampling or pooling layers. A two-layered CNN acting on an image $x \in \mathbb{R}^{H \times W}$ can thus be represented as

$$\mathcal{F}_{CNN}(x) = W_2 \otimes \downarrow_2 (\sigma(W_1 \otimes x + b_1)) + b_2, \quad (3.5)$$

where \downarrow_2 representing downsampling by a factor of 2.

Fully-convolutional neural networks

For tasks pertaining to image-to-image mapping (e.g segmentation, denoising, super-resolution) the fully-convolutional network (FCN) architecture was introduced [12, 13]. An FCN is a combination of an encoder and a decoder, where the encoding happens using convolutional and downsampling layers while the decoding happens using upsampling layers. Borrowing the CNN encoder in Eq. 3.6, we can represent an FCN as

$$\mathcal{F}_{FCN}(x) = W_4 \otimes \uparrow_2 (\sigma(W_3 \otimes \mathcal{F}_{CNN}(x) + b_3)) + b_4, \quad (3.6)$$

where \uparrow_2 representing upsampling by a factor of 2. Upsampling can be obtained using a combination of interpolation and convolution layers or using transposed convolution layers.

3.1.2 Loss

The task T not only decides the choice of the model architecture but also the loss used to gain the experience E . In this section, we will briefly introduce the most common loss functions while highlighting the once employed in this thesis.

Cross-entropy (CE) loss

CE loss is the most commonly used one in machine learning for predicting categorical variables, e.g in the tasks of binary or multi-class classification. It can be denoted as:

$$\mathcal{L} = - \sum_{c \in \text{classes}} y_c \log(\tilde{y}_c), \quad (3.7)$$

where c denotes the class-specific prediction. In probabilistic sense, y_c would be discrete while \tilde{y}_c would be predicted class-wise probability distribution.

CE loss can also be repurposed for tasks such as segmentation and landmark detection. For instance, in case of segmentation, every pixel can be classified into its corresponding class. Similarly, in landmark detection, the feature corresponding to each landmark can be classified into a class corresponding to its location.

Norm-based loss (MSE, MAE)

When the desired prediction (y) is real-valued, norm-based losses such as mean-squared error or mean-absolute error are employed. A p -norm can be used as a loss as shown:

$$\mathcal{L}_p = \|\tilde{y} - y\|_p \quad (3.8)$$

Overlap-based loss

It is common practice to optimise the performance metric P . In tasks of segmentation, soft-Dice [14] is one such example, where the authors optimise a differentiable version of the Dice metric:

$$\mathcal{L}_{dice} = \frac{\sum(y \odot \tilde{y})}{\sum y + \sum \tilde{y}}, \quad (3.9)$$

where \odot denotes point-wise multiplication and the summation happens over all the pixels (or voxels) in the image.

Statistical distance as a loss

Instead of optimising the p -norm distances between the ground truth and prediction, one can also optimise statistical distances. Two such distances used in this thesis are the Kullback-Leibler (KL) divergence and the Earth-mover distance (EMD or Wasserstein metric). Denoting the probability distributions \mathbb{P}_y and $\mathbb{P}_{\tilde{y}}$ defined over the space of probability measures over \mathcal{Y} ,

- The KL divergence is defined as:

$$KL(\mathbb{P}_y \parallel \mathbb{P}_{\tilde{y}}) = \sum_{y \in \mathcal{Y}} \mathbb{P}_y \log \left(\frac{\mathbb{P}_y(y)}{\mathbb{P}_{\tilde{y}}(y)} \right) \quad (3.10)$$

- The Earth-mover distance or Wasserstein metric is defined as:

$$W(\mathbb{P}_y, \mathbb{P}_{\tilde{y}}) = \inf_{\gamma \in \Pi(\mathbb{P}_y, \mathbb{P}_{\tilde{y}})} \mathbb{E}_{(p,q) \sim \gamma} [\|p - q\|], \quad (3.11)$$

where Π denotes the set of all joint distributions γ over p and q whose marginals are \mathbb{P}_y and $\mathbb{P}_{\tilde{y}}$ respectively.

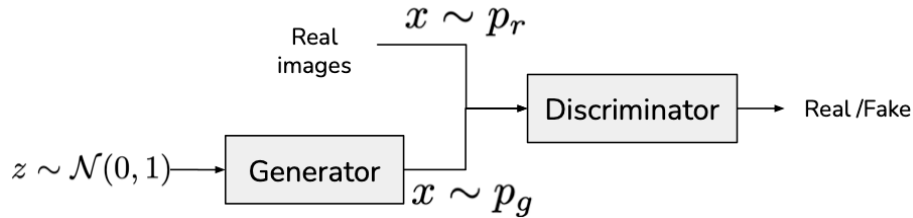


Figure 3.3: Block diagram of a Generative adversarial network.

3.2 Generative models

The previous section largely assumed that a DNN model is used for mapping data, x to its annotation, y , such as a class, segmentation mask etc. Such models are called *discriminative models*, and learn the conditional probability, $P(\mathcal{Y}/\mathcal{X} = x)$. There exist another class of models, called *generative models*, that learn the joint probability distribution of the data and the labels, $P(\mathcal{X}, \mathcal{Y})$, or the data distribution itself, $P(\mathcal{X})$. Such models are predominantly used for generation of new data, e.g. images. However, learning the data distribution or its probability density $p(x)$ is hard. For instance, considering the density functions latent variable decomposition, $p(x) = \int p(x/z)p(z)dz$, summing over all possible latents (z) is intractable. Depending on the approach taken to work around this intractability, there exists a variety of model-families such as generative adversarial networks (GANs), variational autoencoders (VAEs), flow-based models [15], and diffusion models [16]. In this thesis, we focus on GANs and VAEs.

3.2.1 Generative Adversarial Networks

A GAN [17] consists of two models, A discriminator D and a generator G , solving a min-max problem (cf. Fig. 3.3). Specifically, the discriminator estimates the probability of a given input to be a real image or a fake, generated image. On the other hand, the generator takes a noise latent variable z as input to generate or synthesise real-looking data or images. This results in the following optimisation problem:

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x \sim p(x)}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]. \quad (3.12)$$

The discriminator is trained to produce a high probability for a real sample $x \sim p(x)$, obtained by minimising $\mathbb{E}_{x \sim p(x)}[\log D(x)]$. It is also trained to predict a low

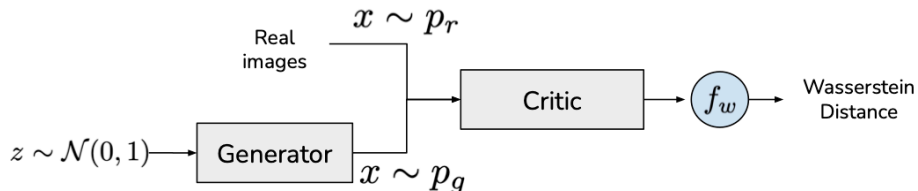


Figure 3.4: Block diagram of a Wasserstein GAN.

probability when a fake sample $G(z)$, $z \sim p(z)$ is fed, by minimising $\mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$. At the same time, G maximises the last term in order to generate real-looking samples.

Ad-hoc optimisation of Eq. 3.12 is known to face several problems such as unstable training, vanishing gradient, mode collapse (where G generates data with high fidelity but low diversity), etc. Several alternatives have thus been proposed to the vanilla GAN such as Wasserstein GAN (WGAN), least-squared GAN (LSGAN), energy-based GAN (EBGAN), boundary-equilibrium GAN (BEGAN) etc, each alters the optimisation problem such that the training is stable while avoiding the problems listed above.

Wasserstein GAN

Arjovsky et al. [18, 19] propose to train the GAN using the Wasserstein distance, introduced in Eq.3.11, to reduce the distance between the real distribution of $x \sim p_r$ and generator’s distribution of it, $x \sim p_g$. However, since it is intractable to determine the infimum over all possible joint distributions $\Pi(p_r, p_g)$, the authors propose an alternative metric using the Kantorovic-Rubenstein duality, which can be used as a loss:

$$\mathcal{L}(p_r, p_g) = W(p_r, p_g) = \max_{w \in \mathcal{W}} \mathbb{E}_{x \sim p_r}[f_w(x)] - \mathbb{E}_{z \sim p(z)}[f_w(g_\theta(z))], \quad (3.13)$$

where f_w comes from a family of K -Lipschitz continuous functions $\{f_w\}_{w \in \mathcal{W}}$. In practice, as shown in Fig. 3.4 the discriminator D is used to model f_w , with tricks such as gradient-clipping to maintain Lipschitz continuity. Essentially, D is no longer a ‘critic’ but a ‘helper’ for estimating the Wasserstein distance.

Energy-based GAN

An autoencoder is an encoder-decoder architecture that learns to encode an input image and decode its exact copy. Instead of using a discriminator that provides

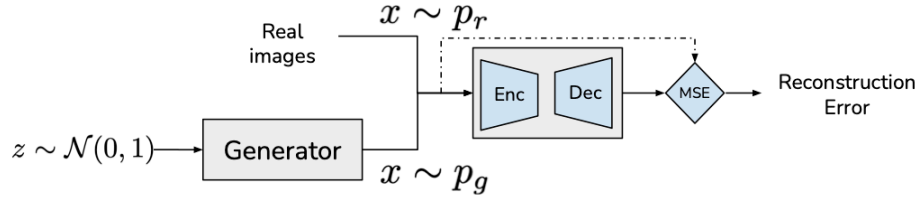


Figure 3.5: Block diagram of an Energy-based GAN.

point gradients based on the input to D being real vs. fake, Zhao et al. [20] design a discriminator that uses an autoencoder that outputs the reconstruction error between the input image and its reconstruction, as shown in Fig. 3.5.

$$D(x) = \|Dec(Enc(x) - x)\|$$

The EBGAN is trained towards two goals: the reconstruction error $D(x)$ should be low for a real input, $x \sim p_r$, and D should be penalised if $D(x)$ is lower than a value (m) for a fake input, $x \sim p_g$. On the other hand, the generator is trained to produce samples resulting in low reconstruction error. These objectives can be represented as:

$$\mathcal{L}_D = D(x) + [m - D(G(z))]^+, \quad (3.14)$$

$$\mathcal{L}_G = D(G(z)), \quad (3.15)$$

where $[\cdot] = \max(0, \cdot)$ denotes the hinge function. To avoid mode collapse, EBGAN also introduced introduces a *pull-away* term as a regulariser acting on the image features S , the output of $Enc(x)$. This is defined as:

$$f_{PT}(S) = \frac{1}{N(N-1)} \sum_{i \in batch} \sum_{j \neq i} \left(\frac{S_i^T S_j}{\|S_i\| \|S_j\|} \right)^2,$$

which is essentially the cosine-distance between the encoded representations of the different images in the batch. In case of a mode-collapse, the cosine distance is maximum, which is discouraged.

3.2.2 Variational Autoencoders

VAEs [21, 22] originate from latent variable modelling of a data distribution $\int p(x/z)p(z)dz$. Since summing over all z is intractable, VAE's narrow the search space by learning

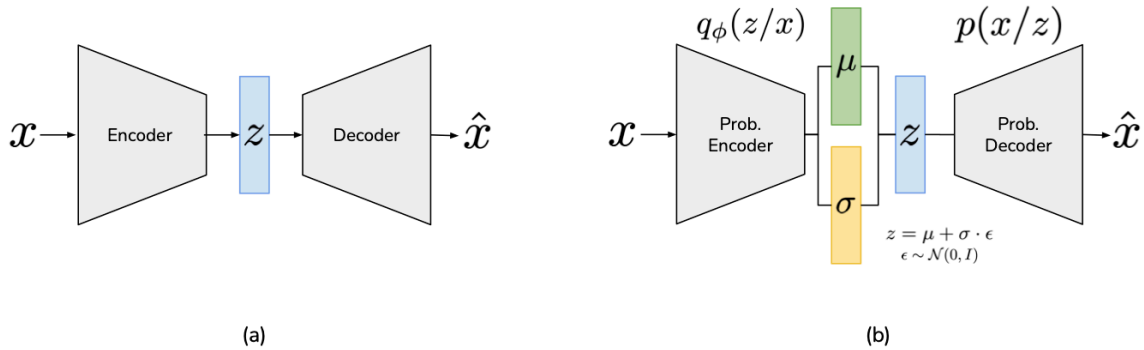


Figure 3.6: Block diagram of an autoencoder and a variational autoencoder.

an encoder $q_\phi(z/x)$, whose role is to output a feasible latent z given an input x . It is necessary that the approximate posterior $q_\phi(z/x)$ is close to the real posterior $p(z/x)$, which forms the basis for deriving the loss function for training the VAE. Specifically, from Eq. 3.10:

$$KL(q_\phi(z/x)||p(z/x)) = \sum_{z \in Z} q_\phi(z/x) \log \frac{q_\phi(z/x)}{p(z/x)} \quad (3.16)$$

$$= \log p(x) + KL(q_\phi(z/x)||p(z)) - \mathbb{E}_{z \sim q_\phi(z/x)} \log p(x/z). \quad (3.17)$$

Rearranging the terms,

$$\log p(x) - KL(q_\phi(z/x)||p(z/x)) = \underbrace{\mathbb{E}_{z \sim q_\phi(z/x)} \log p(x/z)}_{\text{Reconstruction Loss}} - \underbrace{KL(q_\phi(z/x)||p(z))}_{\text{KL divergence}}. \quad (3.18)$$

Incidentally, maximising the left-hand side of this equation is our indirect objective, viz. maximising the log-likelihood of x and minimise the KL-divergence between the approximate and true posteriors. The right-hand side, called the evidence-lower bound (ELBO) of $p(x)$ (because KL divergence in L.H.S is non-negative) forms the loss function for training the VAE. The expectation term requires sampling from q_ϕ , which can be done using the reparameterisation trick, eventually making the VAE trainable using back-propagation. Fig. 3.9 shows an overview of the VAE architecture while comparing it with an autoencoder.

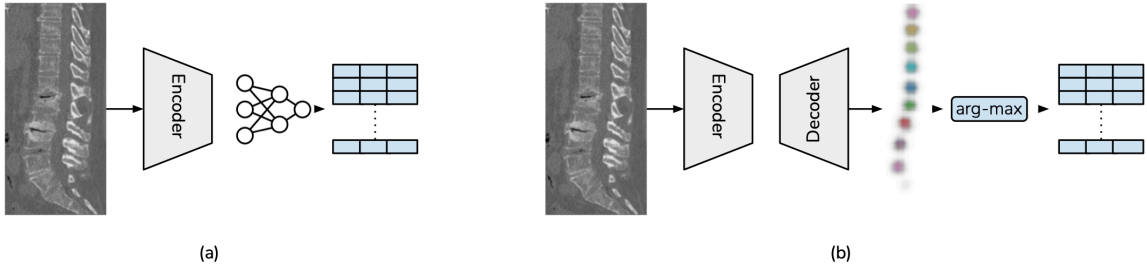


Figure 3.7: Landmark detection: Two traditional approaches towards landmark detection: (a) coordinate-regression (b) heatmap-regression

3.3 Landmark detection and anatomical priors

The objective of landmark detection is to automatically identify landmarks (points-of-interest) in an image or a scan. It is one of the fundamental steps that builds towards other processing tasks such as segmentation, registration, pose-estimation, etc. In the context of medical imaging, landmark detection is also used for critical tasks such as surgery planning, bone age estimation etc. In this section, we introduce the problem of landmark detection in the context of spine and, through prior work, explore how anatomical priors can be incorporated into this task.

Given a 3D scan, $x \in \mathbb{R}^{H \times W \times D}$, the task of landmark detection pertains to detecting the 3D coordinates of N landmarks, $y \in N \times 3$. In vertebrae labelling, N is typically 24, but could be lesser or more depending on the normality of the spine (see Sec. 2.2). There exists two families of approaches, (1) coordinate-regression based methods and (2) heatmap-regression based methods, as shown in Fig. 3.8. Coordinate-regression based approaches directly predict the output coordinates typically using a CNN followed by a dense layer [23, 24, 25, 26]. The task of coordinate-regression inherently involves learning the structural-prior (e.g. in the last dense layer), and shows competitive performance in computer vision tasks such as facial landmark detection and pose-estimation. However, spine images have two problems: not all vertebrae are visible in the image (occlusion) and the image sizes are not standardised (making normalised coordinate predicting infeasible). Therefore, heatmap-based regression methods are preferred for vertebrae labelling [27, 28, 29, 8, 30]. In this, instead of predicting the coordinates directly, the task is modified to predict heatmaps, $y \in \mathbb{R}^{H \times W \times D \times N}$, where every channel $i \in [0, 1, \dots, N - 1]$ contains a Gaussian heatmap at the location of the i^{th} vertebra. Variants of an FCN are employed for this task and are learnt using a regression-based loss such as MSE and

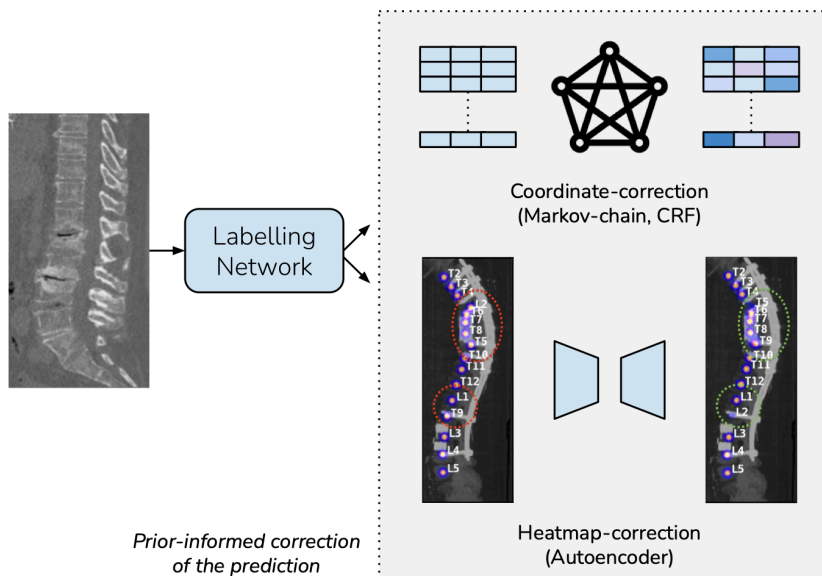


Figure 3.8: Priors in landmark detection: Two traditional approaches towards landmark detection: (a) coordinate-regression (b) heatmap-regression

MAE or a pixel-level cross-entropy loss.

Incorporating anatomical priors

CNNs inherently are local feature extractors and an FCN lacks global feature aggregation. However, efficient vertebrae labelling requires collating local and global priors. The latter can typically be done either post-hoc (predict first, correct anatomy next), wherein a prior-informed corrected happens as a second stage. Post-hoc priors enforcement are traditionally implemented using Markov chains or conditional random fields [31, 32]. Deep neural networks such as denoising auto-encoder trained on ground-truth annotations can also be used to enforce these priors [33, 34, 35]. There exists a second category of prior incorporation, termed *ad-hoc* in which the main model (e.g. FCN) is regularised to make anatomically consistent predictions. An additional loss term can be employed, for instance, a task-specific loss [36] or the reconstruction loss of an autoencoder trained to reconstruct healthy anatomies [37]. This thesis focuses on both *ad-hoc* and *post-hoc* prior enforcement in vertebral labelling.

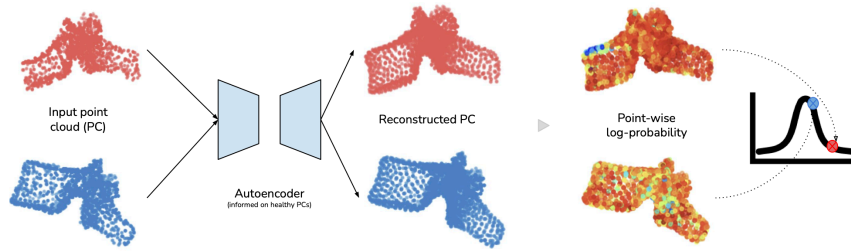


Figure 3.9: An illustration describing outlier-detection in the point-cloud domain

3.4 Point clouds and anomaly detection

Deep learning has predominantly been developed on image and language domains. Architectures such as CNNs and FCNs for images and recurrent neural networks and transformers for language are well established. However, there exists numerous other data modalities which could be optimal for representing data, e.g. graphs for social networks and surface representations for shape models. In the context of this thesis, shape analysis plays a critical role in analysing the spine once it has been processed (e.g. landmarks are detected and vertebrae have been segmented). A point cloud (PC) is a set of N points denoted by $x = \{p_i\}_{i=0}^N$, where p_i denotes a 3D coordinate on the surface. Consider the problem of vertebral fractures, where looking at 3D image patches in high-resolution is compute intensive while lowering the resolution could lead loss of crucial information. *How about representing vertebral shapes as point clouds?* A high-resolution representation with $N = 2048$ would occupy a memory at least two-orders lesser than a 3D image. Point cloud representations have been used for neuroanatomical [38] and vascular tasks [39].

Representation learning and anomaly detection

Consider solving the problem of fracture detection in a data-driven approach. There exists severe data imbalance due to the significantly lesser prevalence of fractured vertebrae compared to healthy ones. Naive, supervised learning approaches such as PointNet [40, 41] fail in this setting. One approach is to use AE-based methods for learning efficient representations for reconstructing healthy vertebrae. Once trained, when an outlier is passed as input, the AE will unsuccessfully project it onto the learnt healthy-representation space resulting in a high reconstruction error. The reconstruction can be represented as $x \rightarrow z \rightarrow \hat{x}$, where z is the the latent represen-

tation and \hat{x} is the reconstruction. In the realm of point clouds, reconstruction error is usually measured using Chamfer distance, computed as:

$$d_ch(x, \hat{x}) = \mathcal{L}_{ae} = \sum_{p \in x} \min_{\hat{p} \in \hat{x}} \|p - \hat{p}\|_2^2 + \sum_{\hat{p} \in \hat{x}} \min_{p \in x} \|p - \hat{p}\|_2^2.$$

Extending AE-based reconstruction to also encode aleoeric uncertainty (data uncertainty), and eventually to a VAE that encodes the latent representations under a Gaussian distribution, classic density-based distance metrics can be employed to detect anomalies [42, 43]. For instance, recalling the notation of the encoder and decoder from Sec. 3.2.2, we can compute the following metrics, for example:

- Reconstruction error for a test sample x_t is defined as,

$$d = \|x_t - \mathbb{E}[p_\theta / (q_\phi(z/x_t))]\|_2^2$$

- Reconstruction likelihood of x_t can be computed as,

$$p = -\log p_\theta(x_t / \mathbb{E}[q_\phi(z/x_t)])$$

Note that corresponding metrics can also be computed in the representation space (z -space) thanks to the Gaussian encoding of the VAE. We refer the reader to [44] work for a list of distance metrics in statistical sense. In this work, we combine the domains of deep learning on point clouds and outlier detection using generative modelling, in order to detect fractures in an unsupervised manner.



Summary of the Contributions

In this section, we will formulate the contributions of this thesis while simultaneously relating them to the preliminaries discussed in Chapter 3 and lay down the organisation of these contributions thesis. Broadly, the thesis contributes to two domains: (1) Vertebrae labelling using anatomical priors and (2) Outlier detection for vertebral fractures; in both cases focusing on efficient computation and explainable implementation geared towards clinical deployment.

Vertebrae labelling using anatomical priors

Recall that labelling vertebra is the task of localising the vertebra and identifying them and an FCN can be used to regress location-based heatmaps. However, due to the high correlation of shape among neighbouring vertebrae, it is important to process enough context in the image for efficient labelling. We propose a novel architecture called Btrfly Net that works with sagittal and coronal reformations of the 3D scan with feature fusion. Working in 2D instead of 3D facilitates the learning of deeper representations. So, *how far can we get in spine labelling when working two dimensions?* Addressing this, a novel architecture, termed BtrflyNet, is proposed to work on orthogonal 2D reformations in **Chapter 5**. Furthermore, an anatomical prior is enforced onto BtrflyNet’s predictions using adversarial training. Specifically, instead of post-hoc correction of predictions, *what if the second stage could be repurposed as a a discriminator, D , in an adversarial regime?* **Chapter 5** also details this approach using a fully-convolutional EBGAN which enforced a local prior (3-5 vertebrae). In **Chapter 6**, the possibility of enforcing a global prior (entire spine) is explored by employing a Wasserstein GAN. Additionally, the question of *how this adversarial prior enforcement affects the latent representations of the Btrfly Net?* is investigated.

Note that in both the chapters above, prior encoding is done implicitly into the Btrfly Net thanks to adversarial training. In other words, the anatomical prior is simultaneously learnt by the discriminator and enforced in a data-driven way. *How can one handle the anatomically abnormal spines mentioned in Sec. 2.2?* This

problem is approached in two phases: First, we adapt a high-resolution network [29] to work in 3D. Second, we employ a prior-informed conditional random field (CRF) conditioned on the likelihood maps of the HRNet to refine the predictions based on five prior-models of the spine. The CRF model being linear makes a real-time solution feasible. **Chapter 7** presents this approach in detail.

Out-of-distribution detection for vertebral fractures

After extracting structure from the image, e.g. landmarks from the image, segmentation from image etc., we look attempt to tackle the challenging task of fracture detection. Recall the efficiency of point-cloud based representation of vertebral shapes and the scarcity of fractured vertebrae. This leads to formulating fracture detection as an outlier detection problem in the PC domain. Specifically, we answer the questions: *How to detect vertebral fractures as outliers with interpretability?* For this, a probabilistic variational autoencoder is proposed, capable of inducing distributions in both the latent space and the reconstruction space, given an input, resulting in human-interpretable fracture detection. This approach is detailed in **Chapter 8**.

PART II

PUBLICATIONS



Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior

Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentinitich, Jan S. Kirschke, & Bjoern H. Menze

Conference: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018.

Synopsis: Robust localisation and identification of vertebrae is essential for automated spine analysis. The contribution of this work to the task is two-fold: (1) Inspired by the human expert, we hypothesise that a sagittal and coronal reformation of the spine contain sufficient information for labelling the vertebrae. Thereby, we propose a butterfly-shaped network architecture (termed Btrfly Net) that efficiently combines the information across reformations. (2) Underpinning the Btrfly net, we present an energy-based adversarial training regime that encodes local spine structure as an anatomical prior into the network, thereby enabling it to achieve state-of-art performance in all standard metrics on a benchmark dataset of 302 scans without any post-processing during inference.

Contributions of thesis author: Conceptualised the project, gathered necessary software resource, developed and implemented the novel architecture and training scheme, lead experimentation and manuscript-writing tasks.

Copyright: Springer Nature AG (Authors permitted to reuse content in full for non-commercial purposes)



Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior

Anjany Sekuboyina^{1,2}(✉), Markus Rempfler¹, Jan Kukačka^{1,2}, Giles Tetteh¹,
Alexander Valentinitzsch², Jan S. Kirschke², and Bjoern H. Menze¹

¹ Department of Informatics, Technical University of Munich, Munich, Germany

² Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany
anjany.sekuboyina@tum.de

Abstract. Robust localisation and identification of vertebrae is essential for automated spine analysis. The contribution of this work to the task is two-fold: (1) Inspired by the human expert, we hypothesise that a sagittal and coronal reformation of the spine contain sufficient information for labelling the vertebrae. Thereby, we propose a butterfly-shaped network architecture (termed Btrfly Net) that efficiently combines the information across reformations. (2) Underpinning the Btrfly net, we present an energy-based adversarial training regime that encodes local spine structure as an anatomical prior into the network, thereby enabling it to achieve state-of-art performance in all standard metrics on a benchmark dataset of 302 scans without any post-processing during inference.

1 Introduction

The localisation and identification of anatomical structures is a significant part of any medical image analysis routine. In spine's context, labelling of vertebrae has immediate diagnostic and modelling significance, e.g.: localised vertebrae are used as markers for detecting kyphosis or scoliosis, vertebral fractures, in surgical planning, or for follow-up analysis tasks such as vertebral segmentation or their bio-mechanical modelling for load analysis.

Vertebrae Labelling. Like several analysis approaches off-late, vertebrae labelling has seen successful utilisation of machine learning. One of the incipient and notable works by Glocker et al. [2], followed by [3] used context-based features with regression forests and Markov models for labelling. In spite of their intuitive motivation, these approaches suffer a setback due to limited FOVs or presence of metal insertions. On a similar footing, [7] proposed a deep multi-layer perceptron using long-range context features. With the emergence of convolutional neural networks (CNN), Chen et al. [1] proposed a joint-CNN as a combination of a random forest for initial candidate selection followed by a

J. S. Kirschke and B. H. Menze—Joint supervising authors.

CNN trained to identify the vertebra based on its appearance and a conditional dependency on its neighbours. Without hand-crafting features this approach performed remarkably well. However, since the CNN works on a limited region around the vertebra, it results in a high variability of the localisation distance. Recently, Yang et al. with [8,9], proposed a deep, volumetric, fully-convolutional 3D network (FCN) called DI2IN with deep-supervision. The output of DI2IN is improved in subsequent stages that employ either message-passing across channels or a convolutional LSTM followed by further tuning with a shape dictionary.

Owing to equivariance of the convolutional operator and limited receptive field, an FCN doesn't always learn the anatomy of the region-of-interest. This is a severe limitation as human-equivalent learning utilises anatomical details aided with prior knowledge. An immediate remedy is to increase the receptive field by going deeper. However, this comes at the cost of higher model complexity or is just unfeasible due to memory constraints when working with volumetric data.

Prior and Adversarial Learning in CNNs. Recent work in [5] and [4] propose encoding (anatomical) segmentation priors into an FCN by learning the shape representation using an auto encoder (AE). The segmentation is expressed in terms of a pre-learned latent space for evaluating a prior-oriented loss, which is then used to guide the FCN into predicting an anatomically sound segmentation. Our approach shares similarities with this approach with certain fundamental differences: (1) Our approach is aimed at localisation, which requires a redefinition of the notion of anatomical *shape*. (2) We employ an AE for shape regularisation, but do not 'pre-train' it to learn the latent space. We train the AE adversarially in tandem with the FCN. Parallels can be drawn between end-to-end learning of priors and learning the distribution of priors using generative adversarial networks (GANs). Both have two networks, a predictor (generator) and an auxiliary network which works on the 'goodness' of the prediction. In medical image analysis where scan sizes are large and data are few, inspired from an energy-based adversarial generation framework (Zhao et al. [11]), it is preferable to employ an adversary providing an anatomically-inspired supervision instead of the usual binary adversarial supervision (vanilla GAN).

Our Contribution. In this work, we propose an end-to-end solution for vertebrae labelling by adversarially training an FCN, thereby encoding the local spine structure into it. More precisely, relying on the sufficiency of information in certain 2D projections of 3D data, we propose: (1) A butterfly-shaped network that operates on 2D sagittal and coronal reformations, combining information across these views at a large receptive field, (2) Encoding the spine's structure into the Btrfly net using an energy-based, fully-convolutional, adversarial auto encoder acting as a discriminator. Our approach attains identification rates above 85% without any post-processing stages, achieving state-of-art performance.

2 Methodology

We present our approach in two stages. First, we describe the Btrfly network tasked with the labelling of the vertebrae. Then we present the adversarial

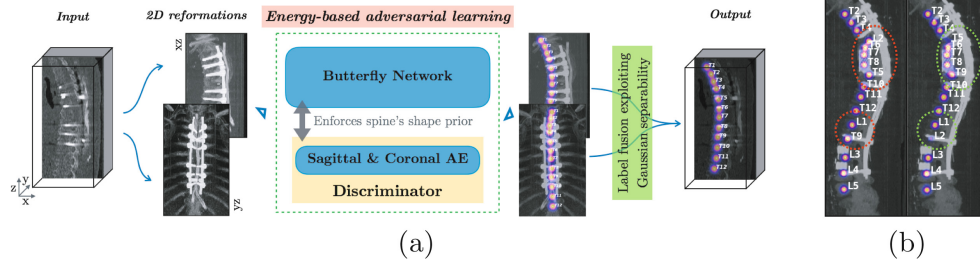


Fig. 1. (a) Overview of our approach. (b) Label correcting capability of the AE when trained as a denoising convolutional auto-encoder (red: corrupted, green: corrected). This motivates the discriminator in our adversarial framework.

learning of the local spine shape with an energy-based auto-encoder acting as the discriminator. Figure 1a gives an overview of the proposed approach and the motivation for prior-encoding is illustrated in Fig. 1b.

2.1 Btrfly Network

Working with 3D volumetric data is computationally restrictive, more so for localisation and identification that rely on a large context so as to capture spatially distant landmarks. Consequently, there is a trade-off between working with low-resolution data or resorting to shallow networks. Therefore, we propose working in 2D with *sufficiently-representative* projections of the volumetric data. The choice of projection is application dependant. Since we are working with bone, we work on sagittal and coronal maximum intensity projections (MIP). The former captures the spine’s curve and the latter captures the rib-vertebrae joints, both of which are crucial markers for labelling. Note that a naive MIP might not always be the optimal choice of projection, eg. in full-body scans where spine is not spatially centred or is obstructed by the ribcage in a MIP. Such cases are handled with a pre-processing stage detecting the occluded spine in the MIP.

Annotations. We formulate the problem of learning the vertebrae labels as a multi-variate regression. The ground-truth annotation $\mathbf{Y} \in \mathbb{R}^{(h \times w \times d \times 25)}$ is a 25-channelled, 3D volume with each channel corresponding to each of the 24 vertebrae (C1 to L5), and one for the background. Each channel i is constructed as a Gaussian heat map of the form $\mathbf{y}_i = e^{-\|x - \mu_i\|^2 / 2\sigma^2}$, $x \in \mathbb{R}^3$ where μ_i is the location of the i^{th} vertebra and σ controls the spread. The background channel is constructed as, $\mathbf{y}_0 = 1 - \max_i(\mathbf{y}_i)$. The sagittal and coronal MIPs of \mathbf{Y} are denoted by $\mathbf{Y}_{\text{sag}} \in \mathbb{R}^{(h \times w \times 25)}$ and $\mathbf{Y}_{\text{cor}} \in \mathbb{R}^{(h \times d \times 25)}$, respectively.

Architecture. We employ an FCN to perform the task of labelling. Since essential information is contained in both the sagittal and coronal reformations, and since the spine is approximately spatially centred in both, fusing this information across views leads to an improved identification. We propose a butterfly-like network (cf. Fig. 2) with two arms (xz- and yz-arms) each concerned with one of the views. The feature maps of both the views are combined after a certain depth in order to learn their inter-dependency.

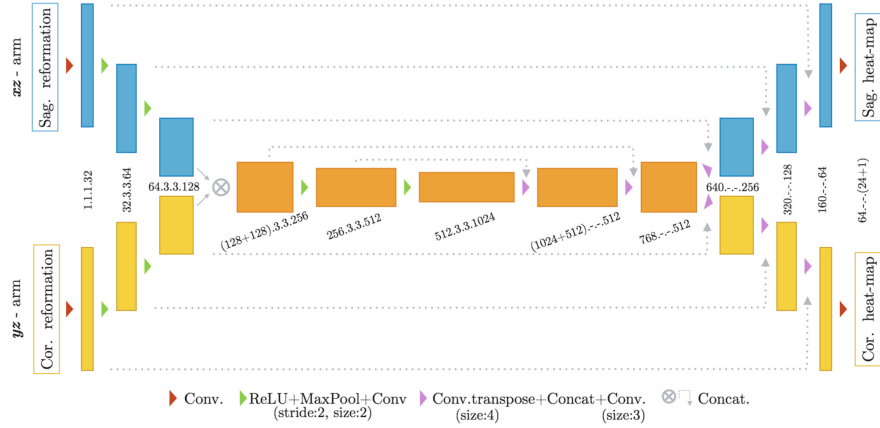


Fig. 2. The Btrfly architecture. The xz- (blue) and the yz-arms (yellow) correspond to the sagittal and coronal views. The kernel’s shape resulting in each of the blocks is indicated as: $\{\text{input channels}\} \cdot \{\text{kern. height}\} \cdot \{\text{kern. width}\} \cdot \{\text{output channels}\}$

Loss. We choose an ℓ_2 distance as the primary loss supported by a cross-entropy loss over the softmax excitation of the ground truth and the prediction. The total loss is expressed as:

$$\mathcal{L}_{b,\text{sag}} = \|\mathbf{Y}_{\text{sag}} - \tilde{\mathbf{Y}}_{\text{sag}}\|^2 + \omega H(\mathbf{Y}_{\text{sag}}^\sigma, \tilde{\mathbf{Y}}_{\text{sag}}^\sigma), \quad (1)$$

where $\tilde{\mathbf{Y}}_{\text{sag}}$ is the prediction of the net’s xz-arm, H is the cross-entropy function, and $\mathbf{Y}_{\text{sag}}^\sigma = \sigma(\mathbf{Y}_{\text{sag}})$, the softmax excitation. ω is the median frequency weighing map (described in [6]), boosting the learning of less frequent classes. The loss for the yz-arm is constructed in a similar fashion and the total loss of the Btrfly net is given by $\mathcal{L}_b = \mathcal{L}_{b,\text{sag}} + \mathcal{L}_{b,\text{cor}}$.

2.2 Energy-Based Adversary for Encoding Prior

Since the Btrfly net is fully-convolutional, its predictions across voxels are independent of each other owing to the spatial invariance of convolutions. Whatever information it encodes is solely due to its receptive field, which may not be anatomically consistent across the image. We propose to impose the anatomical prior of the spine’s shape onto the Btrfly net with *adversarial* learning.

Denoting the projected annotation as \mathbf{Y}_{view} , where $\text{view} \in \{\text{sag}, \text{cor}\}$, a sample annotation consists of a 2D Gaussian at the vertebral location in each channel (except \mathbf{y}_0). Looking at \mathbf{Y}_{view} as a 3D volume enables us in learning the spread of Gaussians across channels and consequently the vertebral labels. However, owing to the extreme variability of FOVs and scan sizes, it is preferable to learn the spread of the vertebrae in parts. Therefore, we employ a fully-convolutional, 3D auto encoder (AE) with a receptive field covering a part of the spine at a time. The absence of fully-connected layers in the AE also removes the necessity to resize the data, making it end-to-end trainable with the Btrfly net. Figure 3a shows the arrangement of the AEs as adversaries w.r.t the Btrfly

net. In an adversarial framework, the Btrfly net acts as the generator (G), and the local manifolds learnt from \mathbf{Y}_{view} influence $\tilde{\mathbf{Y}}_{\text{view}}$ and vice versa.

Discriminator. We devise the 3D adversary (D , cf. Fig. 3b) consisting of the AE as a functional predicting the ℓ_2 distance between the input \mathbf{Y}_{view} and its reconstruction by the AE, $rec(\mathbf{Y}_{\text{view}})$: $D(\mathbf{Y}_{\text{view}}) = E = \|\mathbf{Y}_{\text{view}} - rec(\mathbf{Y}_{\text{view}})\|^2$. This energy, E is fed back into G for adversarial supervision, as in [11]. As it is an energy-based functional, we interchangeably refer to the discriminator as EB- D . Since \mathbf{Y}_{view} consists of Gaussians, it is less informative than an image. Therefore, we avoid using max-pooling by resorting to average pooling. In order to have a receptive field covering multiple vertebrae without using pooling operations, we employ spatially dilated convolution kernels [10] of size $(5 \times 5 \times 5)$ with a dilation rate of 2 (only in image plane), resulting in a receptive field of 76×76 pixels. At 1 mm isotropic resolution, this covers 2 to 3 vertebrae in the lumbar region and more elsewhere.

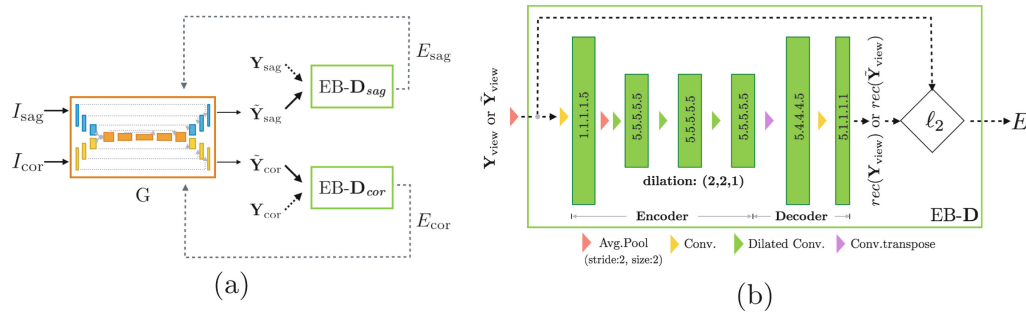


Fig. 3. (a) A overview of adversarial training showing the input to, and the energy-based supervision signal from, the discriminators. (b) The composition of the energy-based discriminator (EB- D). It gives the ℓ_2 reconstruction error as output.

Losses. As in any adversarial setup, EB- D is shown real ($\mathbf{Y}_x (\equiv \mathbf{Y}_{\text{view}})$) and generated annotations ($\mathbf{Y}_g (\equiv \tilde{\mathbf{Y}}_{\text{view}})$), and it learns to discriminate between both by predicting a low E for real annotations, while G learns to generate annotations that would trick D . For a given positive, scalar margin m , the following generator and discriminator losses are optimised:

$$\mathcal{L}_D = D(\mathbf{Y}_x) + \max(0, m - D(\mathbf{Y}_g)), \text{ and} \tag{2}$$

$$\mathcal{L}_G = D(\mathbf{Y}_g) + \mathcal{L}_{b,\text{view}}. \tag{3}$$

The joint optimisation of (2) and (3) for both the EB- D s results in a G that performs vertebrae labelling while respecting the spatial distribution of the vertebrae across channels. We refer to this prior-encoded G as the ‘Btrfly_{pe}’ net.

2.3 Inference

Once trained, an inference for a given input scan of size $(h \times w \times d)$ proceeds as: the desired sagittal and coronal MIP reformations are obtained and given as input to

the xz- and yz-arms of the Btrfly net, resulting in a $(h \times w \times 25)$ sagittal heatmap and $(h \times d \times 25)$ coronal heatmap. The values below a threshold (T , selected on validation set) are ignored in order to remove noisy predictions. As the Gaussian kernel is separable, an outer product of the predictions results in the final heatmap as $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\text{sag}} \otimes \tilde{\mathbf{Y}}_{\text{cor}}$, where \otimes denotes the outer product. The 3D location of the vertebral centroids are obtained as the maxima in their corresponding channels. Note that the EB- D is no longer required during inference as its role in encoding the prior ends with the convergence of the Btrfly_{pe} net.

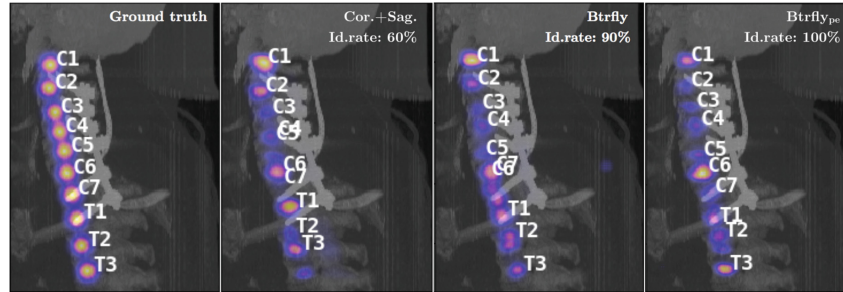


Fig. 4. Effect of prior encoding: the prior-encoded Btrfly_{pe} net successfully performs its task of prevent overlapping labels (C6 & C7), consequently reordering all the vertebral labels. The reported id. rates are per volume.

3 Experiments

The evaluation is performed using a dataset introduced in [3] with a total of 302 CT scans (242 for training and 60 for testing) including various challenges such as scoliotic spines, metal insertions, and highly restrictive FOVs. However, these are cropped to a region around the spine which excludes the ribcage. Thus, a naive sagittal and coronal MIP, without any pre-processing, suffices to obtain the input images for our approach. In order to enhance the net’s robustness, 10 MIPs are obtained from one 3D scan, each time randomly choosing half the slices of interest. This leads to a total of 2420 reformations per view for training (incl. a validation split of 100). We present the experiments with the Btrfly net trained as stand-alone as well as with the prior-encoding discriminator EB- D . Batch-normalisation is used after every convolution layer, along with 20% dropout in the fused layers of Btrfly. Additionally, so as to validate the necessity of the combination of views, we compare the Btrfly net’s performance with that of two networks working individually on the views (denoted as Cor.+Sag. nets). The architecture of each of these networks is similar to one arm of the Btrfly net. The optimiser’s setup in all the three cases is similar: an Adam optimiser is employed with an initial learning rate of $\lambda = 1 \times 10^{-3}$, working on data resampled to a 1 mm isotropic resolution. λ is decayed by a factor of $3/4$ th every 10k iterations to 0.2×10^{-3} . Convergence of all the networks is tested on the validation set.

Evaluation and Discussion. For evaluating the performance of our network with prior work, we use two metrics defined in [2] namely, the *identification rates*

Table 1. Performance comparison of our approach (setting $T = 0$, for a fair comparison) with Glocker et al. [3], Chen et al. [1] and Yang et al. [8]. DI2IN refers to stand-alone FCN, while DI2IN* includes use of message passing and shape dictionary. We do not compare with experiments in [8] that use additional undisclosed data.

Measures	[3]	[1]	DI2IN[8]	DI2IN*[8]	Cor.+Sag.	Btrfly	Btrfly _{pe}
Id.rate	74.0	84.2	76.0	85.0	78.1	81.8	86.1
d_{mean}	13.2	8.8	13.6	8.6	9.3	7.5	7.4
d_{std}	17.8	13.0	37.5	7.8	8.0	5.4	9.3

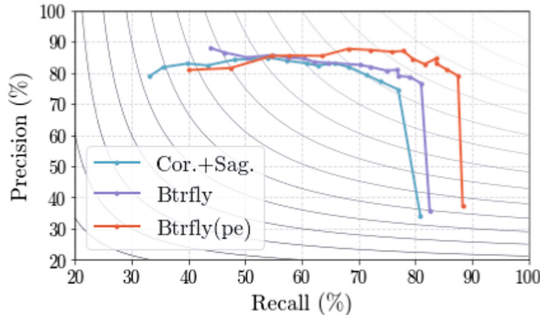


Fig. 5. A precision-recall curve with F1 iso-lines, illustrating the effect of the T during inference. For any T , Btrfly_{pe} offers a better trade-off between P and R .

Table 2. The optimal P and R values based on F1 score, along with the optimal T . R at optimal-F1 of Btrfly_{pe} is comparable to state-of-art.

Approach	P	R	F1
Cor.+Sag. _($T=0.05$)	74.7	77.0	75.8
Btrfly _($T=0.1$)	78.7	79.1	78.9
Btrfly_{pe} ($T=0.2$)	84.6	83.7	84.1

(id. rate, in %) and *localisation distances* (d_{mean} & d_{std} , in mm). We report the measures in Table 1. It lists the performance of three variants of our network and compares them with several recent approaches. We address three main questions through our experiments: (1) *Why the butterfly shape?* Compared to Cor.+Sag. nets, performance improves with the Btrfly net. This is because the combination of views causes the predictions of the Btrfly net to be spatially consistent across views. We also observe a 6% improvement in the id.rate over a naive 3D FCN (DI2IN). (2) *Why the adversarial prior-encoding?* In addition to the advantages of the Btrfly net, the Btrfly_{pe} net possesses adversarially encoded spatial distribution of the vertebrae. This results in about a 4% increase in the id. rate. Compared to the prior work, Btrfly_{pe} net achieves state-of-art measures in both the metrics, and it does so by being a single network trained end-to-end. (cf. Fig. 4) (3) *Relation to latent-space learning?* EB- D is more flexible than the AEs in [4,5] as it learns from scratch and converges to a latent manifold best representing the true as well as generated data. The reconstruction capability of the AE for a generated sample is of interest. Using the output of the AE instead of Btrfly_{pe}, we achieve an id.rate of 75% with a d_{mean} of 19 mm, indicating the AEs' capability of transferring the learning from true to contrastive samples.

Precision and Recall. Localisation distance and id.rate capture the ability of the network in accurately labelling a vertebra. However, both the measures

are agnostic to false positive predictions. Accounting for spurious predictions becomes important especially when dealing with FCNs, as the predictions depend on a locally constrained receptive field. In our case, the false positives are controlled by the threshold T as described in Sect. 2.3. Accounting for these, we define two measures, *precision* (P) and *recall* (R) as: $P = \#hits/\#predicted$ and $R = \#hits/\#actual$, where $\#hits$ is the number of vertebrae satisfying the condition of identification as defined for *id.rate*, $\#predicted$ is the vertebrae in the prediction, and $\#actual$ is the vertebrae actually present in the image. Observe that *id.rate* is measured over all vertebrae in the test set while R is measured *per scan* and averaged over test scans. Figure 5 shows a precision-recall curve generated by varying T between 0 to 0.8 in steps of 0.05, while Table 2 shows the performance at the F1-optimal threshold. In spite of not choosing an recall-optimistic threshold, our networks perform comparably well. Notice the over-arc nature of Btrfly over Cor.+Sag. nets and that of Btrfly_{pe} over others.

4 Conclusions

We validate the sufficiency of 2D orthogonal projections of the spine for localising and identifying the vertebrae by combining information across the projections using a butterfly-like architecture. In addition to looking at a local receptive field like any FCN, our approach considers the local structure of the spine thanks to an adversarial energy-based prior encoding, thereby outperforming the state-of-art approaches as a stand-alone network without any post-processing stages.

Acknowledgements. 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). We acknowledge NVIDIA Corporation’s support with the donation of the Quadro P5000 used for this research.

References

1. Chen, H., et al.: Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In: MICCAI, pp. 515–522 (2015)
2. Glocker, B., et al.: Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In: MICCAI, pp. 590–598 (2012)
3. Glocker, B., et al.: Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In: MICCAI, pp. 262–270 (2013)
4. Oktay, O., et al.: Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. CoRR abs/1705.08302 (2017)
5. Ravishankar, H., et al.: Learning and incorporating shape models for semantic segmentation. In: MICCAI, pp. 203–211 (2017)
6. Roy, A.G., et al.: Error corrective boosting for learning fully convolutional networks with limited data. In: MICCAI, pp. 231–239 (2017)
7. Suzani, A., et al.: Fast automatic vertebrae detection and localization in pathological ct scans - a deep learning approach. In: MICCAI, pp. 678–686 (2015)

8. Yang, D., et al.: Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: IPMI, pp. 633–644 (2017)
9. Yang, D., et al.: Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3D CT volumes. In: MICCAI, pp. 498–506 (2017)
10. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
11. Zhao, J.J., et al.: Energy-based generative adversarial network. CoRR abs/1609.03126 (2016)



Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy

Anjany Sekuboyina, Markus Rempfler, Alexander Valentinitzsch, Jan S. Kirschke, & Bjoern H. Menze.

Journal: Radiology: Artificial Intelligence, 2020

Synopsis:

Purpose.: To use and test a labeling algorithm that operates on two-dimensional reformations, rather than three-dimensional data to locate and identify vertebrae.

Materials and Methods. We improved the Btrfly Net, a fully convolutional network architecture which works on sagittal and coronal maximum intensity projections (MIPs) and augmented it with two additional components: spine localisation and adversarial a priori learning. Furthermore, two variants of adversarial training schemes that incorporated the anatomic, a priori knowledge into the Btrfly Net were explored. The superiority of the proposed approach for labelling vertebrae on three datasets was investigated: a public benchmarking dataset of 302 CT scans and two in-house datasets with a total of 238 CT scans. The Wilcoxon signed rank test was employed to compute the statistical significance of the improvement in performance observed with various architectural components in the approach.

Results. On the public dataset, the proposed approach using the described Btrfly Net with energy-based prior encoding (Btrfly_{pe-eb}) network performed as well

6. LABELING VERTEBRAE WITH TWO-DIMENSIONAL REFORMATIONS OF MULTIDETECTOR CT IMAGES: AN ADVERSARIAL APPROACH FOR INCORPORATING PRIOR KNOWLEDGE OF SPINE ANATOMY ~~as current state of the art methods, achieving a statistically significant ($p < .001$)~~ vertebrae identification rate of $88.5\% \pm 0.2$ (standard deviation) and localization distances of less than $7mm$. On the in-house datasets that had a higher interscan data variability, an identification rate of $85.1\% \pm 1.2$ was obtained.

Conclusion. An identification performance comparable to existing three-dimensional approaches was achieved when labelling vertebrae on two-dimensional MIPs. The performance was further improved using the proposed adversarial training regimen that effectively enforced local spine a priori knowledge during training. Spine localisation increased the generalizability of our approach by homogenising the content in the MIPs .

Contributions of thesis author: Conceptualised the project, gathered necessary software resource, gathered and prepared the in-house data, developed and implemented the novel architecture and training schemes, lead experimentation and manuscript-writing tasks.

Copyright: RSNA (Authors permitted to reuse content in full for non-commercial purposes)


Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy

Anjany Sekuboyina, ME • Markus Rempfler, PhD • Alexander Valentinitzsch, PhD • Bjoern H. Menze, PhD • Jan S. Kirschke, MD

From the Department of Informatics (A.S., B.H.M.) and Department of Neuroradiology, School of Medicine (A.S., A.V., J.S.K.), Technical University of Munich; Department of Diagnostic and Interventional Neuroradiology, Klinikum Rechts der Isar, Ismaninger Str 22, 81675 Munich, Germany (A.S.); and Friedrich Miescher Institute for Biomedical Engineering, Basel, Switzerland (M.R.). Received May 10, 2019; revision requested July 8; revision received January 2, 2020; accepted January 14. Address correspondence to A.S. (e-mail: anjany.sekuboyina@tum.de).

Work supported by the European Research Council under the European Union's Horizon 2020 Research and Innovation program (GA637164-iBack-ERC-2014-STG).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(2):e190074 • <https://doi.org/10.1148/ryai.2020190074> • Content codes: 

Purpose: To use and test a labeling algorithm that operates on two-dimensional reformations, rather than three-dimensional data to locate and identify vertebrae.

Materials and Methods: The authors improved the Btrfly Net, a fully convolutional network architecture described by Sekuboyina et al, which works on sagittal and coronal maximum intensity projections (MIPs) and augmented it with two additional components: spine localization and adversarial a priori learning. Furthermore, two variants of adversarial training schemes that incorporated the anatomic a priori knowledge into the Btrfly Net were explored. The superiority of the proposed approach for labeling vertebrae on three datasets was investigated: a public benchmarking dataset of 302 CT scans and two in-house datasets with a total of 238 CT scans. The Wilcoxon signed rank test was employed to compute the statistical significance of the improvement in performance observed with various architectural components in the authors' approach.

Results: On the public dataset, the authors' approach using the described Btrfly Net with energy-based prior encoding (Btrfly_{pe-eb}) network performed as well as current state-of-the-art methods, achieving a statistically significant ($P < .001$) vertebrae identification rate of $88.5\% \pm 0.2$ (standard deviation) and localization distances of less than 7 mm. On the in-house datasets that had a higher interscan data variability, an identification rate of $85.1\% \pm 1.2$ was obtained.

Conclusion: An identification performance comparable to existing three-dimensional approaches was achieved when labeling vertebrae on two-dimensional MIPs. The performance was further improved using the proposed adversarial training regimen that effectively enforced local spine a priori knowledge during training. Spine localization increased the generalizability of our approach by homogenizing the content in the MIPs.

Supplemental material is available for this article.

© RSNA, 2020

Spine CT is a commonly performed imaging procedure. In this study, we focused on labeling the vertebrae, which is the task of both locating and identifying the cervical, thoracic, lumbar, and sacral vertebrae in a regular spine CT scan. Labeling the vertebrae has immediate diagnostic consequences. Vertebral landmarks help identify scoliosis, pathologic lordosis, and kyphosis, for example. From a modeling perspective, labeling simplifies the downstream tasks of intervertebral disk segmentation and vertebral segmentation.

Previous approaches to labeling fell into one of two broad categories: the traditional model-based approaches and the relatively recent learning-based approaches. Model-based approaches such as those of Schmidt et al (1), Klinder et al (2), and Ma and Lu (3) used a priori information on the spine structure, such as statistical shape models or atlases. Due to their extensive reliance on a priori information, the generalizability of these approaches was limited. From a machine learning perspective, approaches

have existed, ranging from regression forest models working on context features in Glocker et al (4), Glocker et al (5), and Suzani et al (6); a combination of convolutional neural networks and random forest models in Chen et al (7); and three-dimensional (3D) fully convolutional networks in Yang et al (8), followed by recurrent neural networks in Yang et al (9) and Liao et al (10). Most of these approaches work on full 3D multidetector CT scans. Recently, an approach achieved a higher labeling performance by working on two-dimensional (2D) maximum intensity projections (MIPs) using an architecture termed *Btrfly Net*, which was first proposed by Sekuboyina et al (11).

In this study, we improved on the Btrfly architecture and extended it with a spine localization module, thus making the combination more generalizable. Concurrently, inspired by the generative adversarial learning domain, we investigated an a priori learning module that enforced the spine's anatomic a priori knowledge (ie, prior) onto the Btrfly network. Earlier approaches were

Abbreviations

Btrfly_{pe-ch} = Btrfly Net with energy-based prior encoding, Btrfly_{pe-w} = Btrfly Net with Wasserstein distance-based prior encoding, CSI = computational spine imaging, MIP = maximum intensity projection, 3D = three-dimensional, 2D = two-dimensional

Summary

The proposed fully convolutional network architecture, Btrfly Net with energy-based prior encoding, was trained to learn the a priori knowledge of the spine's shape on two-dimensional maximum intensity projections to locate and identify vertebrae at a rate comparable to that of prior methods that operated in three dimensions.

Key Points

- Three-dimensional labeling of vertebrae using two-dimensional sagittal and coronal maximum intensity projections resulted in a computationally lighter but high-performing pipeline when processed using a butterfly-shaped fully convolutional network.
- Employing spine localization as a preprocessing stage enabled the proposed approach to be applicable to scans of any field of view, including complete vertebrae, thus increasing its generalizability to a clinical setting.
- Enforcing anatomic a priori knowledge (in the form of the vertebral arrangement) onto the labeling network using so-called adversarial learning improved the vertebrae identification rate to greater than 88% on a public benchmarking dataset.

aimed at prior learning, such as those of Ravishankar et al (12) and Oktay et al (13). Typically, such approaches consist of two networks, with one primary network solving a task (eg, segmentation) and use of a secondary pretrained network that learns the shape of interest and either “corrects” the primary network’s prediction (12) or “enforces” it on the primary network (13). Our network was similar to these approaches in that it includes two components: a labeling network and an adversary as the secondary network. However, Btrfly Net was fundamentally different in the training and inference processes compared with the previous work (12,13). First, the adversary required no pretraining because it was trained along with the labeling network by penalizing it if the labels deviated from the ground truth distribution. Second, since the adversarial loss was used to update directly the weight of the Btrfly net, the adversary was no longer needed during inference.

Thus, our hypothesis was that by incorporating a spine localization stage before vertebral labeling and then using adversarial learning on an anatomic prior and enforcing it onto the Btrfly network, the labeling performance would be improved while also making the setup generalizable to clinical routines. We describe the use of the prior-encoded Btrfly variant for improved vertebrae identification and labeling compared with the original Btrfly Net. Furthermore, we validate this network for vertebrae labeling on in-house data, supporting its use in a clinical diagnostic setting.

Materials and Methods

Study Datasets

We worked with two datasets: (a) a public benchmarking dataset and (b) a collection of two in-house datasets. The Compu-

tational Spine Imaging (CSI) workshop during the 2014 Medical Image Computing and Computer Assisted Intervention conference by the Department of Radiology at the University of Washington released a public benchmarking dataset, called CSI_{label}. It consists of 302 spine-focused (ie, tightly cropped) CT scans (of which 242 are for training and 60 are for testing) that include fractures, scans with and without contrast material enhancement, abnormal curvatures, and nearly 150 scans comprising postoperative cases with metal implants. The dataset has a mean voxel spacing of approximately $2 \times 0.35 \times 0.35$ mm³ in the craniocaudal \times left-right \times anteroposterior directions and a mean dimension of $(318 \pm 131) \times (172 \pm 24) \times (172 \pm 24)$ at 1-mm³ isotropic resolution. Vertebrae centroids have been manually annotated as described by Glocker et al (5).

Our two in-house datasets consisted of 238 CT scans in total (178 for training and 60 for testing, with fivefold cross validation). It is a collection of examinations of healthy and abnormal spines (eg, osteoporosis, vertebral fractures, degeneration, and scoliosis) of patients 30–80 years old, collected for two previously published retrospective studies (14,15). Ethics approval was obtained from the local ethics committee of the faculty of medicine of the Technical University of Munich for both studies. Because of the retrospective nature of these studies, the need for informed consent was waived. In the study published in 2017, Valentinitich et al (14) established a normative atlas of the thoracolumbar spine in healthy patients using nonenhanced CT scans collected between 2005 and 2014. In the study published in 2019, Valentinitich et al (15) investigated texture analysis techniques for diagnosis of opportunistic osteoporosis using contrast material-enhanced CT scans collected from February 2007 to February 2008. For both studies, cases with metastasis and metal implants were excluded as they would change vertebral texture as well as biomechanical behavior and thus interfere with fracture prediction. These exclusion criteria are based on the older studies and not on the current study we report on here. In this study, we used 65 non-contrast-enhanced scans from Valentinitich et al (14) and 173 contrast-enhanced scans from Valentinitich et al (15). Overall, approximately 20% of the dataset (46 scans) included parts of the rib cage, where 92 patients had fractured vertebrae. Annotations of the vertebrae were automatically derived from available segmentations, performed with an automated algorithm based on shape model matching (2). They were verified by one radiologist (J.S.K.), with more than 15 years of experience in spine imaging, and corrected where necessary.

CT Imaging

All CT scans were performed with either a 256-detector row (Philips Medical Systems, Best, the Netherlands) or a 128-detector row (Siemens Healthineers, Erlangen, Germany) CT scanner and reconstructed using an edge-enhancing kernel. All scans were acquired with 120 kVp and an adaptive tube load. All patients received intravenous contrast material (Iomeron 400; Bracco, Konstanz, Germany) with a delay of 70 seconds, a flow rate of 3 mL/sec, and a body weight-dependent dose between 80 and 100 mL. The mean voxel size of the dataset is approximately $0.7 \times 2.75 \times 0.7$ mm³ in the craniocaudal \times

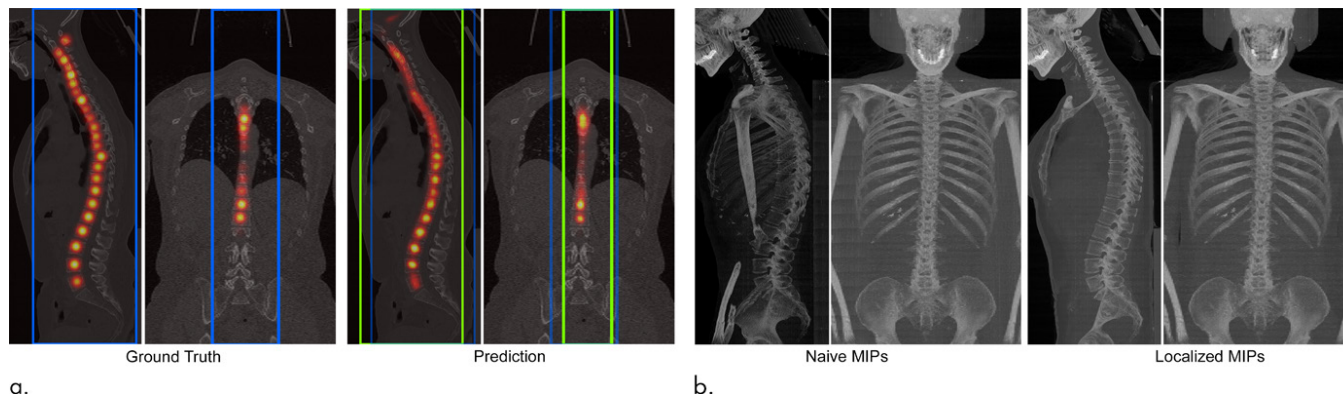


Figure 1: Spine localization and its necessity illustrated on a data sample from the in-house dataset. **(a)** Ground truth and predicted localization maps along with extracted bounding boxes (blue = actual, green = predicted), plotted on sagittal and coronal sections of a CT scan. **(b)** A naive maximum intensity projection (MIP) of a scan shows the rib cage completely occluding the spine (left). This can be handled by extracting a localized MIP of the same scan using the bounding box of the localized spine (right).

left-right \times anteroposterior directions with a mean size of $(565 \pm 111) \times (123 \pm 58) \times (401 \pm 140) \text{ mm}^3$.

Spine Localization as a Preprocessing Stage

For efficiently handling data with diverse fields of view, localizing the spine was an important step that improved the 2D projections on which all subsequent vertebrae labeling stages relied. For localizing the spine, a 3D fully convolutional neural network was employed to regress Gaussian heat maps centered at the vertebrae, followed by extracting a bounding box around the spine. This stage operated at very low resolution (4-mm isotropic), employing a lightweight, fully convolutional U-network. The network architecture is described in Figure E1 (supplement). Figure 1a shows examples of the predicted heat maps and the extracted bounding boxes. With this spine localization stage, we were able to extract a *localized MIP*, which is an MIP across only those sections that contain the spine. Localized MIP projections gave an unoccluded view of the spine (Fig 1b).

Vertebrae Labeling with the Btrfly Net

Networks working in 3D are computationally intensive owing to their features vectors being of $O(N^3)$, where N is one of the scan dimensions. As an alternative to working in a computationally demanding 3D domain, Sekuboyina et al (11) proposed working with 2D sagittal and coronal MIPs by proposing a butterfly-shaped architecture, Btrfly Net. As the Btrfly Net processes 2D data, its feature dimensionality is limited to $O(N^2)$. This reduction in requirement of computational resources allows the design of deeper 2D networks with more convolutional layers leading to higher receptive fields. In this study, we used an improved version of the Btrfly Net. Most importantly, we work with scans at 2-mm isotropic resolution, consequently increasing the receptive field and the representational capacity of the Btrfly Net (Fig E2 [supplement]).

Adversarial Enforcement of the Shape of the Spine onto Btrfly Net

Human anatomy usually follows certain structural rules; thus, anatomic nomenclature has strong tacit assumptions (or pri-

ors). For example, in the spine, vertebra L2 is almost always caudal to L1 and cranial to L3. Enabling a network to learn such priors results in anatomically consistent predictions. In our case, a secondary network or discriminator (D), which was trained along with the primary Btrfly Net, learned the spine's shape and forced the Btrfly Net's predictions to respect this shape. In Sekuboyina et al (11), an energy-based (EB) auto-encoder was used as a discriminator (EB-D). In this study, we improved the architecture of the EB-D and compared it with a purely encoding, Wasserstein distance-based discriminator (W-D). (16). The combination of the Btrfly Net and EB-D was referred to as Btrfly_{pe-eb} (denoting energy-based prior encoding) and that of the Btrfly Net and W-D was referred to as Btrfly_{pe-w} (for Wasserstein distance-based prior encoding). The architectures, the cost functions, and the adversarial training details of both the improved EB-D and proposed W-D are described in Appendix E1 (supplement). Note that EB-D is a fully convolutional network and has a receptive field covering only a fixed part of the spine irrespective of the input scan dimension. On the other hand, W-D has a receptive field covering the full image owing to the dense connections in the architecture. We refer to these two receptive fields as local and global, respectively, and investigate the difference in their behavior.

Inference

Our algorithm included two components—spine localization and vertebrae labeling—as illustrated in Figure 2. Given a multidetector CT scan, the localization stage predicts a heat map (and a bounding box) indicating the location of the spine from which localized MIPs can be extracted and passed to the labeling stage. Each arm of the Btrfly Net labels the sagittal and coronal projections, which are then fused by outer product to obtain the 3D vertebral locations. Note that the role of the discriminators ends with adversarial enforcement of the spine prior onto the Btrfly Network during the training stage and are not a part of the inference path. The improvement in performance due to various components of our approach was assessed using a Wilcoxon signed rank test assuming independence across scans.

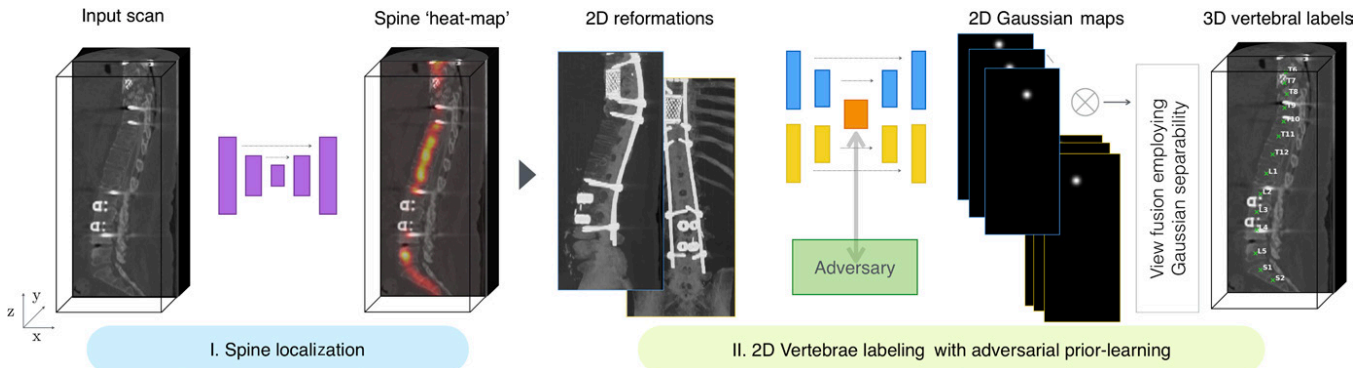


Figure 2: An overview of our labeling approach. Also illustrated is the spine localization stage, which is, in some cases, necessary for our approach to be generalizable to any clinical CT scan including complete vertebrae. 3D = three-dimensional, 2D = two-dimensional.

Performance Evaluation of Spine and Vertebral Detection

Spine localization.—The performance of the localization stage is evaluated with two metrics defined as: (a) intersection over union between the actual and predicted bounding boxes and (b) detection rate, where a detection was successful if the corresponding intersection over union was greater than 50%. It was observed that this overlap suffices for our task of filtering out the obstructions.

Vertebrae labeling.—The labeling performance was evaluated using metrics defined in Glocker et al (4) and Sekuboyina et al (11), namely, identification rate, localization distance, precision, and recall. A vertebra was correctly “identified” if the predicted vertebral location was the closest point to the ground truth location and was less than 20 mm away from it. The distance of this prediction was then recorded as the localization distance. Precision and recall values were used to quantify the relevance of the predictions (eg, accurately labeling all vertebrae in the scan vs labeling a vertebra that was not present in the scan).

Statistical Analysis

Since the test sets are large with 60 scans each for the public and in-house datasets, we employed the nonparametric Wilcoxon signed rank test for validating the statistical significance of the improvement in various performance measures due to our architectural modifications. These modifications included the fusion of the sagittal and coronal views in the Btrfly Net and the inclusion of prior-encoding components in Btrfly_{pe-w} and Btrfly_{pe-cb} Nets.

Results

Successful Spine Localization within the CSI_{label} and In-House Datasets

Localizing the spine at CT is a relatively simple task owing to the higher attenuation (measured in Hounsfield units) of the bone. The network only needs to isolate the spine from the rest of the skeletal structure. Table 1 records the performance of this stage on the two datasets in use. We obtained a mean intersection over union greater than 75% on both datasets. A

Table 1: Performance of the Spine Localization Stage

Measure	CSI _{label} Dataset	In-house Dataset
Mean intersection over union	0.86	0.76
Detection rate (%)	100	96.7

Note.—Intersection over union (ratio between 0.0 and 1.0) is between the bounding boxes around the spine in the ground truth and prediction data. Detection rate is the ratio of successful localizations in the dataset based on the intersection over union. Observe a high detection rate for both the public CSI_{label} dataset and the in-house dataset. CSI = computational spine imaging.

detection rate of 100% on CSI_{label} and 97% on the in-house dataset indicated a successful localization of the spine in all the cases. Note that a complete failure of localization might not be detrimental as it would result only in an occluded MIP. Labeling such a projection could be handled in the subsequent stages with appropriate augmentation if the failure rate is minimal.

Btrfly Network Variants, Btrfly_{pe-cb} and Btrfly_{pe-w} Perform Refined Vertebrae Labeling Compared with the Original Btrfly Network

We evaluated the vertebrae labeling performance in two stages: (a) On the public CSI_{label} dataset, we evaluated the performance of the stand-alone labeling component and compared it with prior work (Tables 2, 3) and (b) on the in-house dataset, we deployed the combination of localization and labeling to evaluate the contribution of localizing the spine (Table 4). To be agnostic to initialization, and since the train-test split is official, we reported the mean performance over three runs of training with independent initializations on the public dataset, while we performed a fivefold cross-validation on the in-house dataset to compensate for any dataset bias. All the improvements in the performance due to our contributions are statistically significant ($P < .05$) according to the Wilcoxon signed rank test.

Contribution of Btrfly architecture.—To validate the importance of combining the views and processing them in the Btrfly

Table 2: Performance Comparison of Our Approach with Prior Work

Approach	Identification Rate (%)				Localization Distance (mm)			
	All	Cervical	Thoracic	Lumbar	All	Cervical	Thoracic	Lumbar
Chen et al (7)	84.2	91.8	76.8	88.1	8.8 ± 13.0	5.1 ± 8.2	11.4 ± 16.5	8.2 ± 8.6
Yang et al (8)	85	92	81	83	8.6 ± 7.8	5.6 ± 4.0	9.2 ± 7.9	11.0 ± 10.8
Liao et al (10)	88.3	95.1	84.0	92.2	6.5 ± 8.6	4.5 ± 4.6*	7.8 ± 10.2	5.6 ± 7.7*
Cor+Sag	85.8 ± 0.8	92.3 ± 0.2*	80.1 ± 2.1	90.0 ± 2.3	6.7 ± 5.4	5.8 ± 5.3	8.2 ± 7.4	7.2 ± 8.1
Btrfly	86.7 ± 0.4	89.4 ± 0.7	83.1 ± 1.0	92.6 ± 1.1	6.3 ± 4.0	6.1 ± 5.4	6.9 ± 5.5	5.7 ± 6.6
Btrfly _{pe-w}	87.7 ± 1.2	89.2 ± 1.3	85.8 ± 1.4	92.9 ± 1.9*	6.4 ± 4.2	5.8 ± 5.4	7.2 ± 5.7	5.6 ± 6.2*
Btrfly _{pe-eb}	88.5 ± 0.2*	89.9 ± 0.2	86.2 ± 0.4*	91.4 ± 1.7	6.2 ± 4.1*	5.9 ± 5.5	6.8 ± 5.9*	5.8 ± 6.6

Note.—Unless otherwise indicated, data are means ± standard deviations. The identification rate is throughout three runs of training with different initializations. Localization distances are the distance computed throughout all vertebrae in the three runs of training (ie, throughout vertebrae in 3 × 60 test set scans). Improvement in the identification rate computed per scan throughout all variants was statistically significant in all runs ($P < .05$). Specifically, the gain in performance of Btrfly_{pe-eb} was statistically significant, with the lowest P value ($< .001$). Btrfly_{pe-eb} = Btrfly Net with energy-based prior encoding, Btrfly_{pe-w} = Btrfly Net with Wasserstein distance-based prior encoding, Cor+Sag = coronal plus sagittal.

* The best performer among the entries.

Net, we compared its performance to a network setup working individually on the coronal and sagittal views without any such fusion of views. This setup was denoted by “Cor+Sag”. The architecture of each of these networks was similar to one arm of the Btrfly Net without the fusion of views. Due to this architectural modification, an increase of 1% in the identification rate was observed between Cor+Sag Net and Btrfly Net, as reported in Table 2.

Effect of adversarial encoding.—Enforcing prior information into the Btrfly Net’s training resulted in a 1% to 2% improvement in the identification rate. We observed that the performance of Btrfly_{pe-eb} was marginally superior to that of Btrfly_{pe-w} (Table 2). This can be explained by the distinction between the locally encoding Btrfly_{pe-eb} and a globally acting Btrfly_{pe-w}. This difference in encoding also explains the high variance of the identification rate of Btrfly_{pe-w} (standard deviation of 0.2

[Btrfly_{pe-eb}] vs 1.2 [Btrfly_{pe-w}]). A qualitative comparison of the three variants in our experiments is shown in Figure 3. The top row shows a use-case with successful labeling, and the bottom row shows an interesting failure use-case where the labeling fails due to the presence of an obstruction along the coronal view.

Table 3: Precision and Recall Analysis

Approach	F1 Optimal Threshold	Precision (%)	Recall (%)	F1
Cor+Sag	0.2	79.5	82.5	79.5
Btrfly	0.33	79.9	85.1	82.4
Btrfly _{pe-w}	0.23	77.8	86.1	81.7
Btrfly _{pe-eb}	0.23	80.2*	87.9*	83.4*

Note.—Reported are the optimal mean precision and recall values based on the F1 score and the mean of the F1-optimal threshold for three runs. Precision = No. of hits/No. predicted and recall = No. of hits/No. of actual, where No. of hits is the number of vertebrae satisfying the condition of identification as defined for identification rate, No. predicted is the vertebrae in the prediction, and No. actual is the vertebrae actually present on the image. Btrfly_{pe-eb} = Btrfly Net with energy-based prior encoding, Btrfly_{pe-w} = Btrfly Net with Wasserstein distance-based prior encoding, Cor+Sag = coronal plus sagittal, No. = number.

* The best performer among the entries.

Table 4: Contribution of Spine Localization

Reformation	Identification Rate (%)				Localization Distance (mm)			
	All	Cervical	Thoracic	Lumbar	All	Cervical	Thoracic	Lumbar
Naive MIP	81.3 ± 3.9	63.1 ± 10.0	79.2 ± 4.6	89.6 ± 4.7	9.5 ± 8.2	14.1 ± 8.4	9.4 ± 7.6	9.5 ± 14.4
Localized MIP	85.1 ± 1.2*	65.7 ± 6.7*	83.7 ± 3.1*	92.0 ± 2.0*	8.1 ± 6.6*	13.4 ± 8.8*	8.1 ± 8.1*	7.7 ± 9.4*

Note.—Data are means ± standard deviations. The in-house dataset contained a significant fraction of scans in which the rib cage causes occluded maximum intensity projections (MIPs). In such cases, localization of the spine enabled extraction of “localized MIPs” from only the spine region. Illustrated in this table is the advantage offered by localized MIPs. Reported measures are obtained using the best-performing architecture (Btrfly_{pe-eb}) from Table 2, averaged with fivefold cross validation.

* The best performer among the entries.

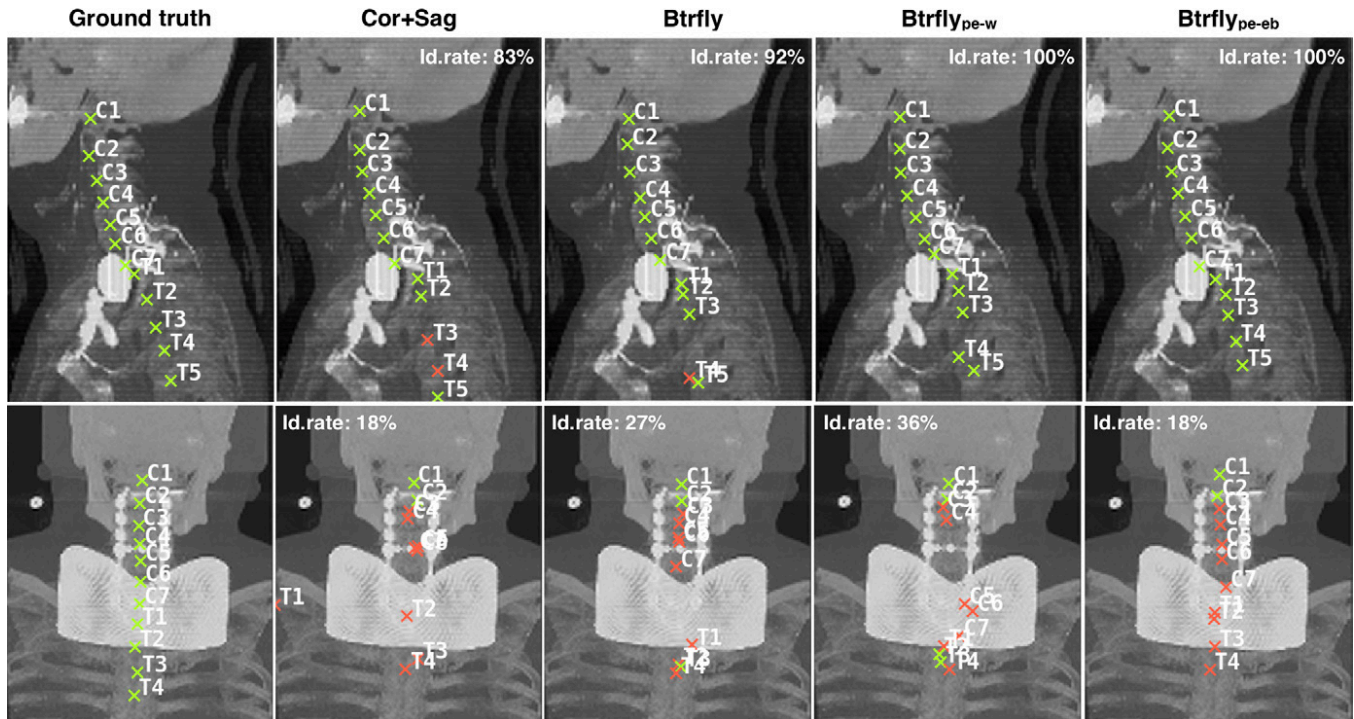


Figure 3: Qualitative comparison of our network architectures shows two cases from the public dataset with one successful labeling (top, sagittal view) and one unsuccessful labeling (bottom, coronal view). The Cor+Sag and Btrfly Nets label mostly in the presence of spatial information. However, the energy-based prior encoding Btrfly Net (Btrfly_{pe-eb}) and Wasserstein distance-based prior encoding Btrfly Net (Btrfly_{pe-w}) hallucinate prospective vertebral labels despite no image information (labels on shaded gray area on coronal view). In addition, Btrfly_{pe-eb} tries to retain the order of vertebral labels. Cor+Sag = coronal plus sagittal. Id. rate = identification rate.

Precision and recall.—A Gaussian image predicted for any vertebra was counted as a final label if the response was higher than a threshold. Thus, the threshold controls the false-positive and true-positive label predictions of the network. Figure 4 shows a precision-recall curve generated by varying the threshold between 0 and 0.8 in steps of 0.1, while Table 3 records the performance at the F1-optimal threshold. Despite not choosing a recall-optimistic threshold, our networks performed comparably well at an optimal-F1 threshold. Notice the overarching nature of Btrfly_{pe-eb} over others at all thresholds.

Rigorous evaluation with respect to identification rate.—As defined in Glocker et al (4), a vertebra was accurately identified if it was the closest to the ground truth and less than 20 mm away. We denote this distance threshold as d_{th} . However, a d_{th} of 20 mm is a weak requirement; for example, in the case of cervical vertebrae, which are quite close to one another, predicting the current vertebra’s landmark on the adjacent vertebra might not be penalized. Demonstrating the spatial precision of our localization, we performed a breakdown test with respect to the identification rates by varying d_{th} between 5 mm and 30 mm in steps of 5 mm. Figure 5 shows the regionwise performance curves obtained for this variation across our setups. Notice the reasonably stable behavior of the curves until $d_{th} = 10$ mm.

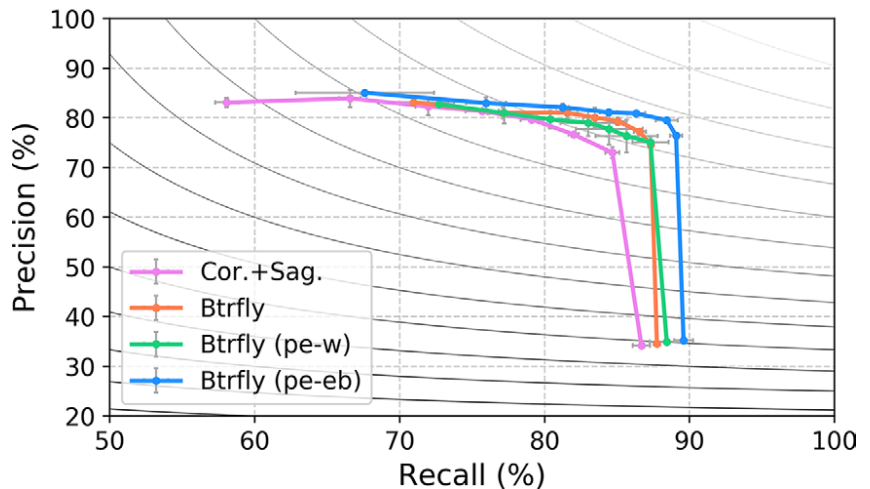


Figure 4: Precision-recall curve with F1 isolines shows the effect of the threshold during inference. For any threshold, Btrfly_{pe-eb} offers a better trade between precision and recall. Cor+Sag = coronal plus sagittal, Btrfly (pe-eb) = Btrfly Net with energy-based prior encoding, Btrfly (pe-w) = Btrfly Net with Wasserstein distance-based prior encoding.

Labeling with improved projections.—The performance of the Btrfly_{pe-eb} architecture (chosen because of its superior performance) on naive MIPs with that of localized MIPs on the in-house dataset is shown in Table 4. Observe an inferior identification rate and high localization distances with naive MIPs due to the lack of visible vertebrae (Fig 6). However, localized MIPs from the spine’s bounding box result in approximately a 4% gain in identification rate.

Discussion

We proposed a generalizable pipeline to localize and identify vertebrae on multidetector CT scans that operates on appropriate 2D projections, unlike prior studies whose authors worked with 3D scans. Specifically, we incorporated an improved version of the Btrfly architecture in combination with a spine localization and a prior-learning stage.

The performance of the localization stage, with intersection over unions greater than 75%, shows that detecting the spine is an easier task. The network localized the spine better

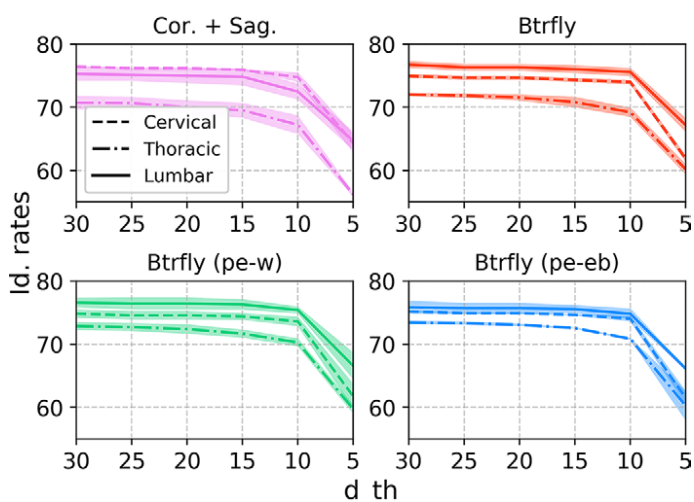


Figure 5: Regionwise variation of identification rates (Id.rates) (y-axis) for different values of the distance threshold (d_{th}) (x-axis) considered as a positive identification. Btrfly (pe-eb) = Btrfly Net with energy-based prior encoding, Btrfly (pe-w) = Btrfly Net with Wasserstein distance-based prior encoding, Cor+Sag = coronal plus sagittal.

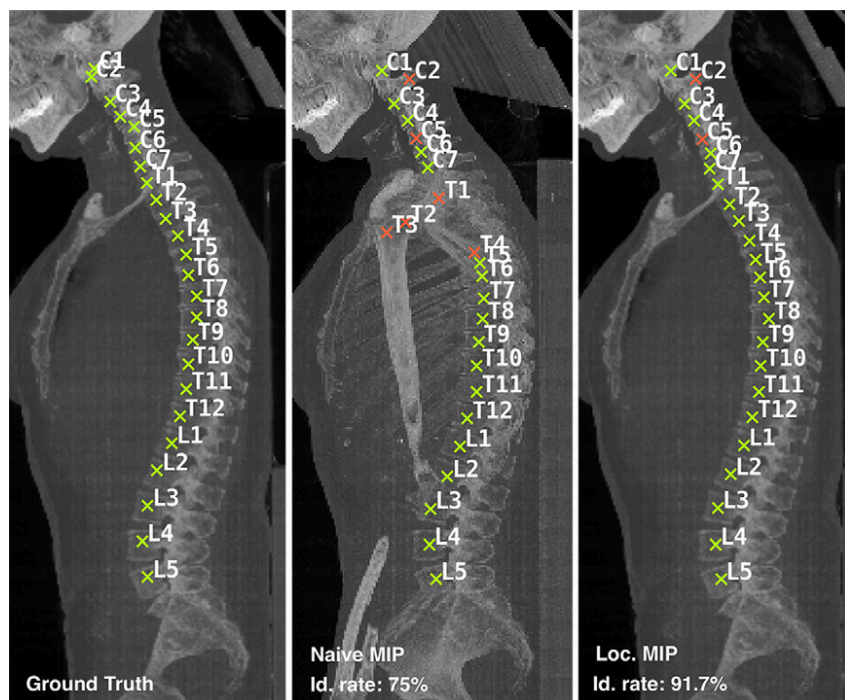


Figure 6: Localized (Loc) maximum intensity projection (MIP) and a naive MIP from the same scan from the in-house dataset using the Btrfly_{pe-eb} network. Observe occlusions in naive MIP and the lack thereof in the localized MIP, resulting in an improved labeling performance. Btrfly (pe-eb) = Btrfly Net with energy-based prior encoding, Id. rate = identification rate.

in the public dataset when compared with the in-house dataset. This can be attributed to the composition of the datasets: CSI_{label} is predominantly composed of thoracolumbar reformations with a uniform axial field of view centered around the spine, while the in-house dataset has a higher variety of scans in terms of axial and coronal fields of view. For the same reason, localizing the spine and extracting MIPs only across the sections containing the spine showed an improvement in labeling performance on the in-house dataset from 81% with naive MIPs to 85% (Table 4) with localized MIPs. Such an improvement was not observed for scans from CSI_{label} due to their relatively uniform fields of view. For the in-house dataset, the improvement can be attributed to the resulting homogeneity in the appearance of the MIPs owing to the localization irrespective of the scan content in the anteroposterior and lateral directions. Such data homogenization also resulted in a more stable learning.

In the module responsible for vertebral labeling, our algorithm contained two principle architectural components: (a) Fusion of the sagittal and coronal MIPs (using the Btrfly net) and (b) incorporation of anatomic prior information using the adversarial discriminators. This architecture, when trained end-to-end, resulted in an identification rate of 88% and localization distances of 6 mm (Table 2) on the public dataset, outperforming prior state-of-the-art methods (7,8,10). The effectiveness of these architectural components can be analyzed in three stages: First, a setup working purely on 2D MIPs (Cor+Sag) readily outperformed the naive 3D fully convolutional network proposed by Yang et al (8).

This was mainly due to the depth of our setup, which can be afforded due to lesser computational load in 2D. Second, fusing the coronal and sagittal views increased the identification rate by approximately 1%. This can be attributed to two reasons: (a) one view now gets access to the key points in the other view, and (b) the combination of views causes the predictions of the Btrfly net to be spatially consistent between views. Third, the Btrfly net's predictions were made to respect an anatomic prior using adversarial discriminators (EB-D or W-D) as a proxy to the postprocessing steps usually employed. For example, Yang et al (8) used a learned dictionary of vertebral centroids to correct a new prediction, and Yang et al (9) and Liao et al (10) used a pretrained recurrent neural network enforcing the vertebrae's sequence to the prediction. Substituting these secondary steps, our Btrfly net was trained adversarially. The second network in our case (the adversary) was neither pretrained nor required during inference. The adversary provided a higher loss when the Btrfly network's prediction did not conform to a normal spine

structure. This enforced the anatomic prior directly onto the Btrfly Net during training. On one of the cross-validation data splits of the in-house dataset, our method achieved a 100% vertebra identification in 34 of the 60 test scans. Moreover, of the remaining 26 scans, seven scans had only one vertebra misidentified and two scans had only two vertebrae misidentified.

Delving deeper into prior learning, EB-D, whose receptive field covered a fixed part of a spine irrespective of the input field of view, enforced the spine's structure locally. In contrast, W-D enforced a global spine prior since the dense layers in it resulted in the receptive field covering the entire input scan. Note that adversarial learning in any form improved the identification rate by 1% to 2%. However, Btrfly_{pe-w} marginally outperformed Btrfly_{pe-eb}. This can be attributed to the fact that a local prior is easier to learn than a global one. Since spine multidetector CT scans have highly varying fields of view, accurately learning a global prior was nontrivial for W-D. In the case of EB-D, the receptive field always covered approximately three vertebrae in the lumbar region (and more elsewhere). Since these subregions are relatively similar across scans, it was relatively easier for EB-D to learn and enforce a local prior. This can also be observed in a lower standard deviation in the identification rate of Btrfly_{pe-eb} across initializations, indicating a more stable performance.

This study has certain limitations stemming from design choices that need further investigation. First, concerning the proposed approach, the optimality of MIPs for labeling requires further study to ascertain cases in which they would break down. From the evaluations performed in this study, it seems to accurately label scoliotic spines, spines with metallic insertions, and spines with fractures, and so forth. However, an organized analysis is lacking. Second, the independent training of the localization stage weakens our approach's "end-to-end trainable" feature. Moreover, the labeling performance depends on the accuracy of localization. When trained until convergence, localization seems to be accurate enough to aid the subsequent labeling stage. However, the robustness of the labeling stage to errors in spine localization is yet to be ascertained. Finally, our study is retrospective, and the computations have been performed on graphical processing units, where a forward pass takes less than 10 seconds. However, even if our model is lighter than its contemporaries, it is demanding for a regular CPU and takes approximately 1 minute per scan. Hence, an effort toward further making the setup lighter is of interest. Finally, consistent with the labels of CSI_{label}, we only considered 24 labels for C1-L5 and did not account for segmentation anomalies, such as L6, or transitional vertebrae, such as a lumbarized S1 vertebra.

Our study presents a simple, efficient algorithm for labeling vertebrae by arguing for processing the spine scans in two dimensions using an appropriate network architecture. Localizing the spine before labeling and forcing the labeling network to respect the spine's anatomic shape during prediction further improved the labeling performance. The entire setup is computationally lighter than its counterparts in the literature, which is a step toward real-time clinical deployment.

Acknowledgments: This study is supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation program (GA637164-iBack-ERC-2014-STG). We also acknowledge support from NVIDIA, with the donation of the Quadro P5000 used for this research.

Author contributions: Guarantors of integrity of entire study, A.S., J.S.K.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; clinical studies, J.S.K.; experimental studies, A.S., M.R., A.V.; statistical analysis, A.S., M.R., A.V., B.H.M.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: A.S. Activities related to the present article: grant/grants pending and travel support from the European Research Council relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. M.R. Activities related to the present article: German Excellence Initiative (and the European Union Seventh Framework Program under grant agreement n291763). Activities not related to the present article: currently employed by the Friedrich Miescher Institute for Biomedical Research. Other relationships: disclosed no relevant relationships. A.V. disclosed no relevant relationships. B.H.M. disclosed no relevant relationships. J.S.K. Activities related to the present article: grant from European Research Council StG iBack and Nvidia. Activities not related to the present article: payment for lectures from Philips Healthcare. Other relationships: disclosed no relevant relationships.

References

- Schmidt S, Kappes J, Bergholdt M, et al. Spine detection and labeling using a parts-based graphical model. In: Information Processing in Medical Imaging. Berlin, Germany: Springer, 2007.
- Klinder T, Ostermann J, Ehm M, Franz A, Kneser R, Lorenz C. Automated model-based vertebra detection, identification, and segmentation in CT images. *Med Image Anal* 2009;13(3):471–482.
- Ma J, Lu L. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. *Comput Vis Image Underst* 2013;117(9):1072–1083.
- Glocker B, Feulner J, Criminisi A, Haynor DR, Konukoglu E. Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2012. https://doi.org/10.1007/978-3-642-33454-2_73.
- Glocker B, Zikic D, Konukoglu E, Haynor DR, Criminisi A. Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2013.
- Suzani A, Seitel A, Liu Y, Fels S, Rohling RN, Abolmaesumi P. Fast automatic vertebrae detection and localization in pathological CT scans: a deep learning approach. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2015.
- Chen H, Shen C, Qin J, et al. Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2015.
- Yang D, Xiong T, Xu D, et al. Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: *Information Processing in Medical Imaging*. Berlin Germany: Springer, 2017.
- Yang D, Xiong T, Xu D, et al. Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3D CT volumes. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin Germany: Springer, 2017.
- Liao H, Mesfin A, Luo J. Joint vertebrae identification and localization in spinal CT images by combining short- and long-range contextual information. *IEEE Trans Med Imaging* 2018;37(5):1266–1275.
- Sekuboyina A, Rempfler M, Kukačka J, et al. Btrfly net: vertebrae labelling with energy-based adversarial learning of local spine prior. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2018.
- Ravishankar H, Venkataramani R, Thiruvankadam S, Sudhakar P, Vaidya V. Learning and incorporating shape models for semantic segmentation. In: *Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2017.

13. Oktay O, Ferrante E, Kamnitsas K, et al. Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. ArXiv 1705.08302 [preprint] <https://arxiv.org/abs/1705.08302>. Posted May 22, 2017. Accessed March 9, 2020.
14. Valentinitich A, Trebeschi S, Alarcón E, et al. Regional analysis of age-related local bone loss in the spine of a healthy population using 3D voxel-based modeling. *Bone* 2017;103:233–240.
15. Valentinitich A, Trebeschi S, Kaesmacher J, et al. Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. *Osteoporos Int* 2019;30(6):1275–1285.
16. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. ArXiv 1701.07875 [preprint] <https://arxiv.org/abs/1701.07875>. Posted January 6, 2017. Accessed March 9, 2020.



Pushing the limits of an FCN and a CRF towards near-ideal vertebrae labelling

Anjany Sekuboyina*, Jannik Irmair*, Suprosanna Shit, Jan S. Kirschke, Bjoern Andres, & Bjoern H. Menze (* denotes equal contribution)

Conference: IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023

Synopsis: In this work, we propose a simple pipeline for labelling vertebrae in a spine CT image composed of a fully convolutional neural network (FCN) and a conditional random field (CRF). Firstly, we adapt the high-resolution network to work on three-dimensional spine CT images and train them with recent advances in deep learning to regress spatial likelihood maps of the vertebral locations. This sets a strong baseline performance for fully automated identification, resulting in a performance comparable to prior state-of-art. Secondly, we employ a prior-informed CRF conditioned on the predicted likelihood maps of the HRNet, thus refining the location predictions. Our custom FCN-CRF solution produces state-of-the-art results in automated labelling tasks for three benchmark datasets achieving identification rates higher than 97%. Finally, we design an interaction module to perform drag-and-drop correction on the CRF output graph. This semi-automated solution achieves near-100% identification with minimal interaction (measured in actions per scan)¹.

Contributions of thesis author: Conceptualised the project, developed and implemented the neural network baseline, prepared all the datasets, shared responsibility for experimentation and manuscript-writing..

Copyright: IEEE (Authors permitted to reuse content in full for non-commercial

¹Code for this work is published at <https://github.com/JannikIrmair/interactive-fcn-crf>

7. PUSHING THE LIMITS OF AN FCN AND A CRF TOWARDS NEAR-IDEAL
VERTEBRAE LABELLING

purposes)

PUSHING THE LIMITS OF AN FCN AND A CRF TOWARDS NEAR-IDEAL VERTEBRAE LABELLING

Anjany Sekuboyina^{*1,2}

Jannik Irmait^{*3}
Bjoern Andres³

Suprosanna Shit^{1,2}
Bjoern Menze¹

Jan Kirschke²

¹ University of Zurich, Switzerland

² Technical University of Munich, Germany

³ Technical University of Dresden, Germany

ABSTRACT

In this work, we propose a simple pipeline for labelling vertebrae in a spine CT image composed of a fully convolutional neural network (FCN) and a conditional random field (CRF). Firstly, we adapt the high-resolution network to work on three-dimensional spine CT images and train them with recent advances in deep learning to regress spatial likelihood maps of the vertebral locations. This sets a strong baseline performance for fully automated identification, resulting in a performance comparable to prior state-of-art. Secondly, we employ a prior-informed CRF conditioned on the predicted likelihood maps of the HRNet, thus refining the location predictions. Our custom FCN-CRF solution produces state-of-the-art results in automated labelling tasks for three benchmark datasets achieving identification rates higher than 97%. Finally, we design an interaction module to perform drag-and-drop correction on the CRF output graph. This semi-automated solution achieves near-100% identification with minimal interaction (measured in actions per scan). Code for this work is published at <https://github.com/JannikIrmait/interactive-fcn-crf>.

Index Terms— landmark detection, spine, vertebrae, fully convolutional neural network, conditional random fields

1. INTRODUCTION

Benefits of computer-aided diagnosis have been demonstrated in multiple instances [1, 2, 3]. In the case of the spine, such automated diagnosis or processing typically requires accurate identification of the vertebrae [4, 5], which enables the success of downstream tasks such as segmentation, surgery planning, automated reporting, etc. With the prevalence of deep learning, automated processing has seen rapid growth. However, the clinical adoption of such a system remains limited [6], with one argument being that such systems are difficult to integrate into clinical workflows, especially because they will work as support-tools.

^{*}Equal contribution.

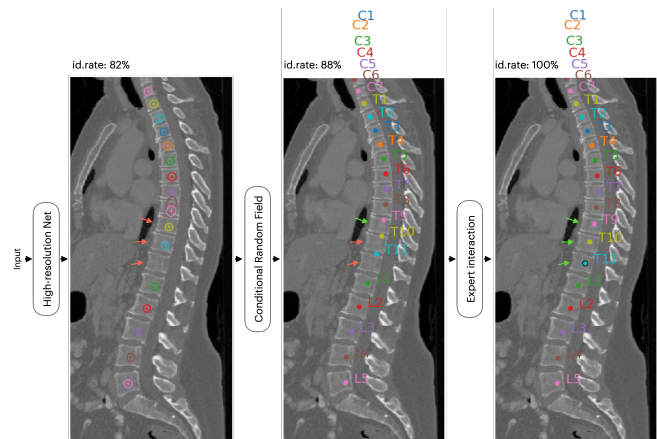


Fig. 1: An overview of the proposed method composed of a fully-convolutional neural network, a prior-informing conditional random field, and an interaction module facilitating high gain with minimal interaction (in this case with missing T12, only T11 is corrected and the other vertebrae also move while respecting the anatomy).

In this paper, we work toward the objective of automatically labelling vertebrae with a focus on the clinical adaptability of such a system, wherein an expert wishes to interact as little as possible, if at all, with any automated system. To this end, we specifically propose a vertebrae labelling method consisting of three components: The first component consists of a fully-convolutional neural network (FCN) predicting the spatial likelihoods (heatmaps) of the vertebrae. In spite of demonstrating strong identification performances, the convolutional network's prediction is still 'local'. The second component consists of a conditional random field (CRF), whose role is to incorporate 'global' information into the prediction. This is done by considering the joint distribution of both the localisation predicted by the FCN as well as the relative positioning of neighbouring vertebrae that encode the spine shape prior. In our pipeline, the CRF increases the number of perfectly identified scans (wherein every vertebra is correctly identified and localised) to more than 90%. We observe that the remaining errors occur due to abnormal spines

(e.g., severe vertebral fracture, transitional vertebrae such as T13 or L6, etc.) and argue that any simple yet automated system will typically find it challenging to learn such abnormal cases reliably. One can think of increasing model complexity to counter these edge cases, however, the limited amount of train data and no explicit quality control would add little-to-no values to their reliable deployability. Therefore, as the third component, we opt for a human-in-loop approach for correcting such erroneous predictions. Importantly, thanks to the anatomical prior enforced by the CRF, our pipeline requires very little interaction. This is because adjusting one landmark also refines the location of the other landmarks, thanks to the CRF in the backend, while staying faithful to the FCN predictions (see Fig. 1). For instance, on the CSI dataset [7], we achieve a near-perfect identification rate of 98% with just two mouse-actions over the entire test set.

Related work. Sekuboyina et al. [8], provide a holistic overview of the prior work tackling vertebrae labelling till 2021. A common theme among the recent best performing, deep learning-based methods involves supporting the *local* feature extraction of the FCN with a module that enforces the *global* spine prior. The latter is done using generative learning [9], recurrent neural networks [10], or an auxiliary FCN [11]. Recently, attention-based models, i.e., Spine Transformers [12] have shown very high vertebrae labelling performance. Keeping the model deployability in mind, in this work, we side-step from intricate methodological modelling, and instead focus on simplicity. Of interest is a pioneering works by Glocker et al. [7] and Chen et al. [13], which use a regression forest or an FCN to obtain candidate locations of the vertebrae and refine it using a hidden Markov model (HMM) on the class-wise votes. Our composite of a learnable model and a graphical model is similar to these approaches in principle. However, in contrast to these approaches, we propose an FCN to learn the calibrated spatial likelihood (instead of any voting mechanism) and fit a conditional random field on these likelihoods, transiting from the pixel domain to the 3D coordinate domain (Further details in Sec. 2).

2. METHODS

2.1. HRNet3D for heatmap-based vertebrae labelling

Our first objective is to predict for each voxel $v \in H \times W \times D$ and every vertebra $i \in C$ a likelihood $Y_i^v \in [0, 1]$ of voxel v being part of the vertebra i . Specifically, we aim to predict $Y \in \mathbb{R}^{H \times W \times D \times |C|}$, where Y_i is a likelihood map of the i^{th} vertebra and is defined by $e^{-\|x-x_i\|^2/2\sigma^2}$. Here, x , x_i , and σ denote the voxel location, the vertebra’s location, and the user-defined spread of the likelihood, respectively.

Architecture. For this, we adapt the HRNet [14], an FCN proposed for facial landmark detection¹. HRNet is a collection of parallel convolution streams performing feature ex-

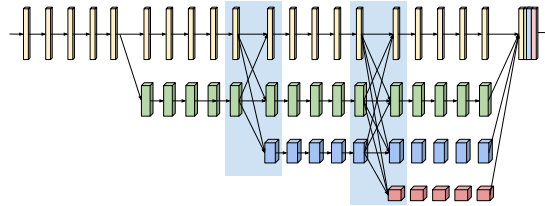


Fig. 2: A schematic of the HRNet3D illustrating the parallel convolution streams and multi-resolution fusion.

traction at multiple resolutions with multi-scale fusion. After every stage, a lower-resolution convolution stream is added. In our case, a 3D input image is convolved to $\frac{1}{4}^{th}$ of its resolution using two strided convolution kernels. Following this, we employ four stages of parallel convolution streams as shown in Fig. 2. The last stage has resolution streams of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. The final prediction is a concatenation of the multiple resolutions upsampled using trilinear interpolation to the image dimension.

Loss. We propose to use a combination of the mean-squared error (MSE) and cross-entropy (CE) losses to train the network. We observed that the CE-loss was essential for convergence in a low-data regime while the MSE-loss resulted in calibrated heatmaps, which is essential for the following CRF stage. Specifically, we first construct a background channel $Y_0 = 1 - \text{sum}_C(Y)$, a summation along the C -dimension and concatenate it to the likelihood map. We now learn to predict $Y^+ = [Y_0; Y] \in \mathbb{R}^{H \times W \times D \times (1+|C|)}$ using the following loss:

$$L = \|Y^+ - \text{softmax}(\hat{Y})\|^2 + H(Y^+, \hat{Y}) , \quad (1)$$

where H denotes the CE function. Observe that the prediction \hat{Y} contains logits, whose probability-normalised (softmax) values are calibrated to the ground truth.

2.2. Gaussian CRF as a spine-prior model

Whereas the HRNet predicts the location of every single vertebra independently, a conditional random field (CRF) [15] can take into account the relative position of the vertebrae to each other. More specifically, the relative positioning of the vertebrae is informed by a *spine-prior model* that captures the shape of the spine. For an example, see Fig. 3a.

Let C be the set of vertebrae and let $G = (C, E)$ be a graph where the set of edges E are the pairs of vertebrae whose relative positions are considered. In this work we consider only the relative positions of neighbouring vertebrae, i.e. G is a path with $E = \{\{i, i+1\} \mid i \in \{1, \dots, |C|-1\}\}$. The Gaussian CRF with respect to G is defined as the joint probability distribution

$$\mathbb{P}(x \mid \hat{Y}) = \prod_{i \in C} \phi_i(x_i \mid \hat{Y}_i) \prod_{ij \in E} \psi_{ij}(x_i - x_j) , \quad (2)$$

¹<https://github.com/HRNet/HRNet-Facial-Landmark-Detection>

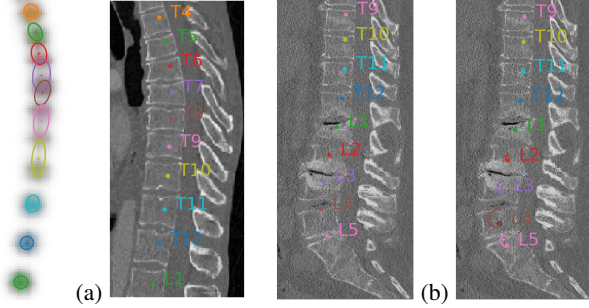


Fig. 3: (a) **CRF.** The heatmaps (left) predicted by the HRNet are ambiguous for vertebrae T7 to T10. This is also reflected in the fitted unary distributions, which are depicted by the ellipses. Thanks to the *spine-prior*, the MAP estimate (right) of the CRF is capable of resolving this ambiguity. (b) **Interaction.** Depicted on the left is an erroneous MAP estimate which was caused by several fractures. Depicted on the right is the path (brown) that a user dragged the vertebra L4 to correct its location and the CRF automatically readjusts all other vertebral locations.

where ϕ_i is the unary distribution of vertebra i which is inferred from the corresponding heatmap \hat{Y}_i and ψ_{ij} is the distribution of the relative position of vertebrae i and j which encode the spine shape prior. We model both ϕ and ψ as three dimensional Gaussian distributions with mean $\mu \in \mathbb{R}^3$ and positive-definite covariance matrix $\Sigma \in \mathbb{R}^{3 \times 3}$. The parameters μ_{ij} and Σ_{ij} of ψ_{ij} are maximum likelihood estimates from the training set, while the parameters μ_i and Σ_i of ϕ_i are fitted to the predicted heatmap \hat{Y}_i of a test image.

The joint probability density (2) of the CRF is a $3 \cdot |C|$ -dimensional Gaussian distribution with mean m and precision matrix P , where

$$P = \begin{pmatrix} P_{11} & \dots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \dots & P_{nn} \end{pmatrix}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \quad Pm = b$$

with

$$P_{ii} = \Sigma_i^{-1} + \sum_{j \in C: ij \in E} \Sigma_{ij}^{-1} \quad \text{for } i \in C$$

$$P_{ij} = \begin{cases} -\Sigma_{ij}^{-1} & \text{for } ij \in E \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i, j \in C, i \neq j$$

$$b_i = \Sigma_i^{-1} \mu_i + \sum_{j \in C: ij \in E} \Sigma_{ij}^{-1} \mu_{ij} \quad \text{for } i \in C.$$

The maximum a posteriori (MAP) estimate of (2) is precisely the joint mean m , the prior-adjusted prediction, and it can be efficiently computed by solving the linear system $Pm = b$.

2.3. Expert interaction

Abnormal spines are unlikely by definition and unlikely also w.r.t. the shape prior of our CRF. Thus, we cannot expect this

CRF to locate all vertebrae of abnormal spines correctly. To address this, we have implemented an application that allows for an intuitive and efficient correction of these erroneous predictions by an expert. The expert interactions consist of selecting a misplaced vertebra and dragging it to the correct location. This corrected location is used as *evidence* in the CRF, and the locations of the remaining vertebrae are updated with respect to that *evidence*. An advantage of using a Gaussian CRF (compared to a more involved graphical model) is that the MAP estimate for the given *evidence* can be computed readily by solving a linear system of $3 \cdot |C|$ equations and $3 \cdot |C|$ unknowns and the user can observe in real-time how changing the location of one vertebrae effects the locations of the other vertebrae (see Fig. 3b).

3. RESULTS & DISCUSSION

Implementation. The HRNet3D was trained on isotropic spine CT images resampled to 2mm. In the case of the VERSE datasets, the scans were cropped to the spine's extent using a U-Net operating at 5mm [8]. The network was trained with a batch size 2. An initial learning rate of 1e-3 was used and annealed using cosine annealing till 1e-5 with warm restarts every 1500 batches. The hyperparameters were tuned on 20% of the train data, and the final model trained on all train data used the same hyperparameters.

Before fitting the parameters of the unary terms ϕ of the CRF to the predicted heatmaps \hat{Y} , we set all heatmap values below 0.1 to zero to suppress noise. Further, we set the precision Σ_i^{-1} of all vertebrae i for which the maximal value in the corresponding heatmap \hat{Y}_i is below 0.3 to zero such that the MAP estimate of vertebrae with highly uncertain HRNet prediction is informed only by the prior model. The thresholds 0.1 and 0.3 are not tuned for the individual datasets.

We benchmark our approach on three datasets: CSI-2014 [7], VERSE'19 and VERSE'20 [8]. CSI 2014 consists of 242 CT images for training and 60 images for testing. VERSE'19 and '20, have two testing phases: public and hidden. Their splits among train/public-test(I)/hidden-test(II) are 80/40/40 and 113/103/103 images respectively.

In line with prior work [7], we measure the performance using the identification rate (id-rate) and mean localisation distance (d_{mean}) for CSI-2014. Please note the slightly different scan-level computation of id-rate in the VERSE datasets [8]. We note that the id-rate was computed as required by the dataset. Additionally, we introduce a new measure, perf-id, that captures the amount of human intervention needed post-automated labelling. perf-id (perfect-identification) is computed as the ratio of the number of scans that obtain an id-rate of 100% to the total number of scans in the dataset.

A strong FCN baseline. Table. 1 compares our approach with prior state-of-arts for every datasets. Observe that HR-Net3D, a single-stage FCN performs competitively even with

the multi-staged methods. For instance, on CSI-2014, Payer et al. [11] uses a two-stream FCN followed by hard-coded post-processing and Chen et al. [13] uses an FCN followed by a HMM. Similarly, on the VERSE datasets, HRNet’s performance is comparable to the top performers, which are also multi-staged. For instance, Chen et al. (VERSE’19) use a heavily engineered labelling scheme. Chen et al. (VERSE’20) re-label every vertebra after segmenting it. Our HRNet3D, however, is learnt end-to-end.

Table 1: Performance comparison on the three datasets with prior work. d_{mean} in millimeters. **Blue** values indicate the best in a fully-automated, no-oracle setting while the **bold** values indicate the best overall. * indicates an oracle for the prior selection.

	Methods	id-rate	d_{mean}	perf-id
CSI-2014	Payer et al. [11]	96.0%	2.9	–
	Chen et al. [13]	94.7%	2.6	–
	Tao et al. [12]	92.2%	4.8	–
	HRNet3D	96.3%	2.5	88.3%
	HRNet3D + CRF	97.3%	2.3	93.3%
	HRNet3D + CRF* ... +1-click interaction	97.3%	2.3	93.3%
VERSE’19 - I	Chen et al. [8]	96.9%	4.4	–
	Payer et al. [8]	95.6%	4.3	–
	Tao et al. [12]	97.2%	4.3	–
	HRNet3D	94.5%	2.7	75.0%
	HRNet3D + CRF	97.9%	2.8	92.5%
	HRNet3D + CRF* ... +1-click interaction	98.2%	2.8	95.0%
VERSE’19 - II	Chen et al. [8]	86.7%	7.1	–
	Payer et al. [8]	94.3%	4.8	–
	Tao et al. [12]	96.7%	4.8	–
	HRNet3D	94.4%	2.8	75.0%
	HRNet3D + CRF	95.2%	2.8	85.0%
	HRNet3D + CRF* ... +1-click interaction	95.4%	2.8	87.5%
VERSE’20 - I	Chen et al. [8]	95.6%	2.0	–
	Payer et al. [8]	95.1%	2.9	–
	HRNet3D	94.1%	2.6	80.6%
	HRNet3D + CRF	93.3%	2.8	77.7%
	HRNet3D + CRF* ... +1-click interaction	94.1%	2.8	80.6%
	HRNet3D + CRF* ... +1-click interaction	97.1%	2.6	92.2%
VERSE’20 - II	Chen et al. [8]	96.6%	1.4	–
	Payer et al. [8]	92.8%	2.9	–
	HRNet3D	96.4%	2.6	87.4%
	HRNet3D + CRF	96.8%	2.4	90.3%
	HRNet3D + CRF* ... +1-click interaction	97.1%	2.4	92.2%
	HRNet3D + CRF* ... +1-click interaction	98.0%	2.4	97.1%

Fixing mislabelling with the CRF. Observe that the perf-id for the HRNet3D leaves considerable room for improvement for a fully-automated setting. We identify that the failed cases typically are due to the FCN being uncertain locally (see Fig 3). The role of the CRF is to incorporate a global, anatomical prior into HRNet3D’s predictions. In Table. 1, we observe that such prior information results in a consistent im-

provement of not only the id-rate but also the perf-id. However, it is important to note that one *spine-prior model* cannot account for all spine anatomies, e.g the abnormal anatomies with transitional vertebrae (T13), L6, or a missing T12 or L5. Hence, we collect five *prior* models for the spine: **1.** the normal spine {C1-C7, T1-T12, L1-L5, S1, S2}, **2.** the normal spine with an additional T13, **3.** the normal spine without T12, **4.** the normal spine with an additional L6, and **5.** the normal spine without L5.

Now, how do we choose which model to enforce? We propose to reuse the channel-responses in the predictions for prior model selection (e.g. if the channel corresponding to T13 has a response higher than 0.5, we choose model **2** from above). This is reported as CRF in Table 1. **Till this point our method is fully automated and outperforms the prior state-of-art on CSI-2014, VERSE’19-I and VERSE’20-II.** However, one can design other auxiliary models to inform the prior selection; for instance the shape classification model in Chen et al. (VERSE’20, [8]). Using the ground truth as the oracle, we report the maximum performance that can thus be achieved using a CRF denoted as CRF* in Table 1.

The minimal-interaction setup. We limit the allowed interaction to reflect the little time available for such corrections as part of the clinical routine. Specifically, we allowed only ONE action per image, and this mouse-action is allowed ONLY if the expert knows that the action leads to perfect identification (here, we use the oracle to infer which prior model to use as the expert is assumed to be able to always identify the correct model). Such minimal-interaction allowance lets us validate the convenience offered by the CRF-based prior models, which is a key objective of this work. Table 1 shows the immense gain obtained even in these very restricted interactive settings. The final id-rates are above 97% for every dataset, with perf-id going above 96%. We believe that a sophisticated prior-model selection will offer further advantage, including for VERSE’20-I. We also note that a comprehensive study with multiple experts is part of the future work. Interestingly, VERSE’19 gets a 100% identification and CSI-2014 gets a near-100% (only two failed cases due to missing vertebral bodies).

4. CONCLUSIONS

We introduced a deep neural network for automatically labelling all vertebrae in a spine CT and a Gaussian CRF with a shape prior for refining the labelling. The identification rate of the FCN-CRF composition is state-of-the-art on three datasets and is improved further by single drag-and-drop expert interactions, especially for abnormal spines. Our CRF module readjusts other, relevant vertebrae from single expert interaction in real-time by solving a linear system and thereby has the potential for implementation in medical hardware and clinical routine.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

This research work was funded by the Helmut Horten Foundation, Switzerland. A. Sekuboyina and J. Kirschke are managing directors of bonescreen GmbH.

7. REFERENCES

- [1] Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, et al., “Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs,” *Radiology*, vol. 294, no. 1, pp. 199–209, 2020.
- [2] Philipp Kickingereeder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, et al., “Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study,” *The Lancet Oncology*, vol. 20, no. 5, pp. 728–740, 2019.
- [3] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, et al., “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.
- [4] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, et al., “A vertebral segmentation dataset with fracture grading,” *Radiology: Artificial Intelligence*, vol. 2, no. 4, 2020.
- [5] Malek Hussein, Anjany Sekuboyina, Maximilian Loeffler, Fernando Navarro, Bjoern H Menze, and Jan S Kirschke, “Grading loss: a fracture grade-based metric loss for vertebral fracture detection,” in *MICCAI*. Springer, 2020, pp. 733–742.
- [6] Thomas Davenport and Ravi Kalakota, “The potential for artificial intelligence in healthcare,” *Future healthcare journal*, vol. 6, no. 2, pp. 94, 2019.
- [7] Ben Glocker, Johannes Feulner, Antonio Criminisi, David R Haynor, and Ender Konukoglu, “Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans,” in *MICCAI*. Springer, 2012, pp. 590–598.
- [8] Anjany Sekuboyina, Malek E Hussein, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, et al., “Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images,” *MedIA*, vol. 73, pp. 102166, 2021.
- [9] Anjany Sekuboyina, Markus Rempfler, Jan Kukačka, Giles Tetteh, Alexander Valentinitich, et al., “Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior,” in *MICCAI*. Springer, 2018, pp. 649–657.
- [10] Haofu Liao, Addisu Mesfin, and Jiebo Luo, “Joint vertebrae identification and localization in spinal ct images by combining short-and long-range contextual information,” *IEEE TMI*, vol. 37, no. 5, pp. 1266–1275, 2018.
- [11] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler, “Integrating spatial configuration into heatmap regression based cnns for landmark localization,” *MedIA*, vol. 54, pp. 207–219, 2019.
- [12] Rong Tao, Wenyong Liu, and Guoyan Zheng, “Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine cts via 3d transformers,” *MedIA*, vol. 75, pp. 102258, 2022.
- [13] Yizhi Chen, Yunhe Gao, Kang Li, Liang Zhao, and Jun Zhao, “vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model,” *IEEE TMI*, vol. 39, no. 2, pp. 387–399, 2019.
- [14] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, et al., “Deep high-resolution representation learning for visual recognition,” *IEEE TPAMI*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [15] Martin J Wainwright, Michael I Jordan, et al., “Graphical models, exponential families, and variational inference,” *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.



Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis

Anjany Sekuboyina, Markus Rempfler, Alexander Valentinitzsch, Maximilian Loeffler, Jan S. Kirschke, & Bjoern H. Menze

Journal: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019.

Synopsis: We propose an auto-encoding network architecture for point clouds (PC) capable of extracting shape signatures without supervision. Building on this, we (i) design a loss function capable of modelling data variance on PCs which are unstructured, and (ii) regularise the latent space as in a variational auto-encoder, both of which increase the auto-encoders' descriptive capacity while making them probabilistic. Evaluating the reconstruction quality of our architectures, we employ them for detecting vertebral fractures without any supervision. By learning to efficiently reconstruct only healthy vertebrae, fractures are detected as anomalous reconstructions. Evaluating on a dataset containing 1500 vertebrae, we achieve area-under-ROC curve of >75%, without using intensity-based features..

Contributions of thesis author: Conceptualised the project, gathered necessary software resource, gathered and prepared the in-house data, developed and implemented the novel point-cloud-based architecture, lead experimentation and manuscript-writing tasks.

Copyright: Springer Nature AG (Authors permitted to reuse content in full for non-commercial purposes)



Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis

Anjany Sekuboyina^{1,2(✉)}, Markus Rempfler³, Alexander Valentinitzsch²,
Maximilian Loeffler², Jan S. Kirschke², and Bjoern H. Menze¹

¹ Department of Informatics, Technical University of Munich, Munich, Germany

² Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany

anjany.sekuboyina@tum.de

³ Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

Abstract. We propose an auto-encoding network architecture for point clouds (PC) capable of extracting shape signatures without supervision. Building on this, we (i) design a loss function capable of modelling data variance on PCs which are unstructured, and (ii) regularise the latent space as in a variational auto-encoder, both of which increase the auto-encoders' descriptive capacity while making them probabilistic. Evaluating the reconstruction quality of our architectures, we employ them for detecting vertebral fractures without any supervision. By learning to efficiently reconstruct only healthy vertebrae, fractures are detected as anomalous reconstructions. Evaluating on a dataset containing ~1500 vertebrae, we achieve area-under-ROC curve of >75%, without using intensity-based features.

1 Introduction

One of the consequences of the numerous algorithms proposed for segmenting organs, tissues, the spine etc. involves analysing their anatomical shapes, eventually contributing towards population studies [1], disease characterisation [2], survival analysis [3], etc. Employing convolutional neural networks (CNN) for this task involves processing voxelised data due to its Euclidean nature. Such voluminous representation, however, is inefficient, especially when the masks are binary and the *shape information* corresponds to its surface profile. Alternatively, surface meshes (a collection of vertices, edges, and faces) or active contours could be used. Since the data is no longer Euclidean, a conventional CNN is unusable. Graph convolutional networks (GCN) [4] were thus developed by redefining the notion of 'neighbourhood' and 'convolution' for meshes and graphs. However,

J. S. Kirschke and B. H. Menze—Joint supervising authors.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32226-7_42) contains supplementary material, which is available to authorized users.

if the number of nodes is high, GCNs (esp. spectral) become bulky. Moreover, each mesh is treated as a domain, making mesh registration a requisite.

An alternative surface representation is a set of 3D points in space, referred to as the **point clouds** (PC). A PC represents the surface just with a set of N vertices, thus avoiding both the cubic-complexity of voxel-based representations and the $N \times N$ dimensional, sparse, adjacency matrix of meshes. However, despite their representational effectiveness, PCs are permutation invariant and do not describe data on a structured grid, preventing the usage of standard convolution. To this end, we work with an architecture capable of processing PCs (*point-net*, [5]), and design a network capable of reconstructing PCs thereby extracting shape signatures in an unsupervised manner.

Uncertainty and Latent Space Modelling. Unlike supervised learning on PCs [6], we set out to obtain shape signatures from PCs without supervision, building towards a relatively less explored topic of *auto-encoding* point clouds. This involves mapping the PC to a latent vector and reconstructing it back. Since the PCs are unordered, PC-specific reconstruction losses replace traditional ones [7,8]. Extending auto-encoders (AE) based on such a loss, we propose to improve its representational capacity by regularising the latent space to make it compact and by modelling the variance that exists in a PC population. We claim that this results in learning improved shape signatures, validating the claim by employing the extracted features for unsupervised vertebral fracture detection.

Vertebral Fracture Detection. There exists an inherent shape variation in vertebral shapes within the spine of a single patient (e.g. cervical–thoracic–lumbar) along with a natural variation in a vertebra’s shape in a population (e.g. L1 across patients, cf. Fig. 1). Additionally, osteoporotic fractures start without significant shape change and progress into a vertebral collapse. Hence, fracture detection in vertebrae is non-trivial. Added to this, limited availability of fractured vertebrae makes the learning of supervised classifiers non-trivial. In literature, several classification systems exist mainly based on vertebral height measurement [9] or analysing sub-regions of the spine in sagittal slices [10]. However, an explicit shape-based approach seems absent. Evaluating the representational ability of the proposed AE architectures, we seek to analyse vertebral shapes and eventually detect vertebral fractures using the extracted latent shape features.

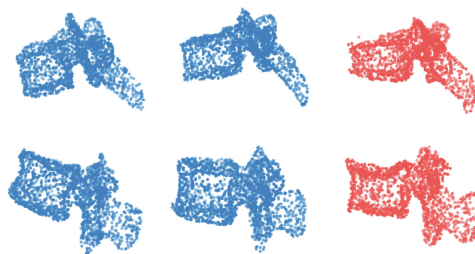


Fig. 1. Variation among vertebral shapes: compare the higher variation between healthy (blue) vertebrae of different classes (T3, top and L1, bottom) w.r.t the relatively lower variation within-class between fractured (red) and healthy vertebrae. (Color figure online)

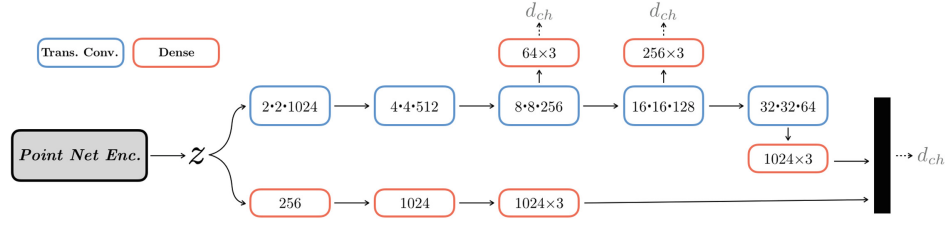


Fig. 2. Point cloud auto-encoder (*pAE*): architectural details of decoding path constructing a point cloud from a latent vector. Top arm is convolutional while bottom arm is fully-connected. Transposed convolution ($- \cdot - \cdot$ channels) have a stride of 2. Since encoder is an adapted *point-net* [5], we detail its architecture in the supplement.

Our Contribution. Summarising the contributions of this work: (1) We build on existing point-net-based architectures to propose a point-cloud auto encoder (*pAE*). (2) Reinforcing this architecture, we incorporate latent space modelling and a more challenging uncertainty quantification. (3) We present a comprehensive analysis of the reconstruction capabilities of our *pAEs* by investigating their utility in detecting vertebral fractures. We work with an in-house, clinical dataset (~ 1500 vertebrae) achieving an area-under-curve (AUC) of $>75\%$ in detecting fractures, even without employing texture or intensity-based features.

2 Methodology

We present this section in two stages: First, we introduce the notation used in this work and describe a point-net-based architecture capable of efficiently auto-encoding point clouds. Second, we build on this architecture to model the natural variance in vertebrae while regularising the latent space.

2.1 Auto-Encoding Point Clouds

Given accurate voxel-wise segmentation of a vertebra, a *point cloud* (PC) can be extracted as a set of N points denoted by $X = \{p_i\}_{i=0}^N$, where p_i represents a point by its 3D coordinate (x_i, y_i, z_i) . Additionally, p_i could also represent other point specific features such as normal, radius of curvature etc. So, each vertebra is represented by a PC of dimension $N \times m$ (in this work, $N = 2048$ vertices and $m = 3$ coordinates, with the vertices randomly subsampled from a higher resolution mesh). Recall the lack of a regular coordinate space associated with the PC and that any permutation of these N points represents the same PC. Thus, a unique variant of deep networks is incorporated for processing PCs.

Architecture. An AE consists of an encoder mapping the PC to the latent vector and a decoder reconstructing the PC back from this latent vector, i.e $X \mapsto z \mapsto X$. As the encoder, we employ a variant of the point-net architecture [5]. The latent vector, z , respects the permutation invariance of the PC

and represents its shape signature. As a decoder, taking cues from [7], we construct a combination of an up-convolutional and dense branches taking z as input and predicting \hat{X} , the reconstructed X . The convolutional path, owing to its neighbourhood processing, models the ‘average’ regions, while the dense path reconstructs the finer structures. This combination of the point-net and the decoder forms our point cloud auto-encoding (p AE, or interchangeably AE) architecture as illustrated in Fig. 2.

Loss. Reconstructing point clouds requires comparing the predicted PC with the actual PC to back-propagate the loss during training. However, owing to the unordered nature of PCs, usual regression losses cannot be employed. Two prominent candidates for such a task are the Chamfer distance and the Earth Mover (EM) distance [7]. We observed that minimising EM distance ignores the natural variation in shapes (e.g. the processes of the vertebrae) and reconstructs only a mean representation (e.g. the vertebral body), as validated in [7]. Since we intend to model the natural variance in the data, using EM distance is undesirable in our case. We thus employ the Chamfer distance computed as:

$$d_{ch}(X, \hat{X}) = \mathcal{L}_{ae} = \sum_{p \in X} \min_{\hat{p} \in \hat{X}} \|p - \hat{p}\|_2^2 + \sum_{\hat{p} \in \hat{X}} \min_{p \in X} \|p - \hat{p}\|_2^2. \quad (1)$$

In essence, d_{ch} is the distance between a point in X and its nearest neighbour in \hat{X} and vice versa.

2.2 Probabilistic Reconstruction

From a generative modelling perspective, an AE can be seen to predict the parameters of Gaussian distribution imposed on X , i.e. $p_{\Theta}(X) = \mathcal{N}(X|\hat{X}, \hat{\Sigma})$, parameterised by the weights of the AE denoted by Θ . Determining the distribution parameters, viz. optimising for the AE weights, now involves maximising the log-likelihood of X , resulting in:

$$\Theta^* = \arg \max_{\Theta} \log p_{\Theta}(X) = \arg \min_{\Theta} \frac{1}{2}(X - \hat{X})^T \hat{\Sigma}^{-1}(X - \hat{X}) + \frac{1}{2} \log |\hat{\Sigma}|. \quad (2)$$

This perspective towards auto-encoding enables us to extend the p AE to encompass the data variance ($\hat{\Sigma}$) while modelling the latent space, as described in following sections. It is important to note that the difference $X - \hat{X}$ is not well defined for point clouds, requiring us to opt for alternatives.

Assuming $\Sigma = \mathbb{I}$, implying an independence among the elements of X and an element-wise unit variance, results in the familiar mean squared error (MSE), $\mathcal{L} = \|X - \hat{X}\|^2$. Based on the parallels between MSE and the Chamfer distance (Eq. 1), we design σ -AE and σ -VAE, as illustrated in Fig. 3.

σ -AE. The assumption of unit covariance, as in AE, is inherently restrictive. However, modelling an unconstrained covariance matrix is infeasible due to quadratic complexity. A practical compromise is the independence assumption. Thus, representing covariance as, $\Sigma = \text{diag}\{\hat{\sigma}_{p_1}^2, \dots, \hat{\sigma}_{p_i}^2, \dots, \hat{\sigma}_{p_N}^2\}$, where

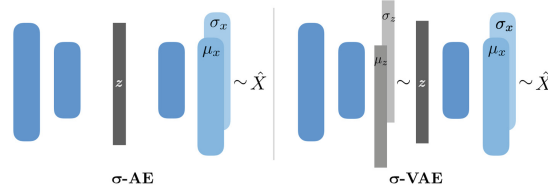


Fig. 3. Probabilistic reconstruction architectures: \sim indicates a sampling operation. Since a point’s variance has a smaller scale compared to its mean, the variance is predicted using a softplus activation (added with $\epsilon = 10^{-6}$ for stabilising divisions) and uses a layer parallel to the one predicting the mean.

$\hat{\sigma}_{p_i}^2$ denotes the variance corresponding to p_i , Eq. (2) morphs to a loss function as:

$$\mathcal{L} = \sum_{\hat{p} \in \hat{X}} \sigma_{\hat{p}}^{-2} \|p_i - \hat{p}_i\|^2 + \log \sigma_{\hat{p}}^2 \quad (3)$$

This optimisation models the aleoteric uncertainty [11]. Equation 3 is an attenuated MSE, where a high variance associated to a point down-weights its contribution to the loss. However, due to the lack of a reference grid in the point cloud space, the notion of uncertainty being associated to a data point (eg. pixel, spatial location etc.) is absent. We propose to associate the notion of variance to every point, \hat{p}_i . This results in the variance-modelling Chamfer distance:

$$\mathcal{L}_{\sigma ae} = \sum_{p \in X} \min_{\hat{p} \in \hat{X}} \sigma_{\hat{p}}^{-2} \|p - \hat{p}\|_2^2 + \sum_{\hat{p} \in \hat{X}} \sigma_{\hat{p}}^{-2} \min_{p \in X} \|p - \hat{p}\|_2^2 + \log \sigma_{\hat{p}}^2 \quad (4)$$

Observe the slight abuse of notation in Eq. 4, wherein the variance at a predicted point, $\sigma_{\hat{p}}$, actually represents the variance of the coordinate elements of p , i.e $\{\sigma_{\hat{x}}, \sigma_{\hat{y}}, \sigma_{\hat{z}}\}$. Current notation is chosen to avoid clutter.

Variational and σ -Variational AE. An alternative approach for modelling $p(X)$ involves modelling its dependency over a latent variable z , which is distributed according to a known prior $p(z)$. A variational auto-encoder (VAE) operates on these principles and involves maximising a lower bound on the log-evidence (referred to as ELBO) of the data described as below:

$$\log p(X) \geq \mathbb{E}_{z \sim q_{\phi}(z|X)} [\log p_{\theta}(X|z)] - \text{KL}[q_{\phi}(z|X) \parallel p_{\theta}(z)], \quad (5)$$

where $q_{\phi}(z|x)$ is the approximate posterior of z learnt by the encoder and parameterised by ϕ . $p_{\theta}(X|z)$ is the data likelihood modelled by the decoder and parameterised by θ . $p_{\theta}(z)$ is the prior on z .

Maximising ELBO is equivalent to maximising the log-likelihood of X while minimising the Kullback-Leibler divergence between the approximate and true prior. Representing the combination as $\mathcal{L}_{rec} + \beta \mathcal{L}_{KL}$, where \mathcal{L}_{rec} is the reconstruction loss seen in earlier sections. β is a scaling factor weighing the contribution of the two losses appropriately. Standard practice assigns Gaussian distributions for $q_{\phi}(z|x) \sim \mathcal{N}(z|\mu_z, \sigma_z)$ and $p(z) \sim \mathcal{N}(z|\mathbf{0}, \mathbf{1})$ (cf. Fig. 3). Thus,

\mathcal{L}_{KL} models the latent space to follow a Gaussian distribution inline with the prior. Incorporating this into the point cloud domain, results in an objective function for a PC-based VAE (or σ -VAE) as $\mathcal{L}_{vae} = \mathcal{L}_{ae/\sigma ae} + \beta \mathcal{L}_{KL}$. Thus, σ -VAE acts as a AE capable of modelling the data variance while regularising the latent space. The prior on the latent space also imparts point cloud generation capabilities to σ -VAE.

2.3 Detecting Fractures as Anomalies

Examining the descriptive ability of our p AE architectures in auto-encoding PCs, we utilise them for detecting vertebral fractures. Assuming the AE is trained only on ‘normal’ patterns, a fracture can be detected as an ‘anomaly’ based on its ‘position’ in latent space. We inspect two measures for this purpose:

1. Reconstruction error or Chamfer distance: AEs trained on healthy samples fail to accurately reconstruct anomalous ones, resulting in a high d_{ch} .
2. Reconstruction probability or likelihood [12]: Expected likelihood $\mathbb{E}[p_{\theta}(X)]$ of an input can be computed for σ – architectures (cf. Eq. 2). For any input PC, X_{in} , it is computed by $\mathcal{N}(X_{in}|\mu_{\theta}, \Sigma_{\theta})$ with the predicted mean and variances. We expect fractured vertebrae to be less *likely* than healthy ones.

Intuitively, relying on the reconstruction error or likelihood for detecting anomalies requires the learnt ‘healthy’ latent space to be representative. Both σ -AE and the VAE work towards this objective. In σ -AE, predictive variance down-weighs the loss due to highly uncertain points in the PC. This suppresses the interference due to natural variation in the vertebral PCs. On the other hand, VAE acts directly on the latent space by modelling the encoding uncertainty ($X \mapsto z$). The σ -VAE encompasses both these features.

Inference. A given vertebral PC is reconstructed and the reconstruction error and (or) likelihood are computed. This vertebra is said to be fractured if the reconstruction error is greater than a threshold, T_{rec} , or its likelihood is lesser than a threshold, T_l . T_{rec} and T_l are determined on the validation set.

3 Experiments and Discussion

We present this section in two parts: first, we explore the auto-encoding, variance modelling, and generative capabilities of our AE networks. Second, we deploy these architecture to detect vertebral fractures without supervision.

Data preparation: We evaluate our architecture on an in-house dataset with accurate voxel-level segmentations converted into PCs. The dataset consists of 1525 healthy and 155 fractured vertebrae, denoted as $(1525H + 155F)$ vertebrae. Since we intend to learn the distribution of healthy vertebrae, we do not use any fractured vertebrae during training. The validation and test sets consists of $(50H + 55F)$ and $(100H + 100F)$ vertebrae, respectively. For the supervised

baselines, the train set needs to contain fractured vertebrae. Thus, validation and test sets were altered to contain $(50H + 55F)$ and $(55H + 55F)$ vertebrae. *Training:* The architecture of the encoder and the decoder is similar across all architectures (cf. Fig. 3) except for the layers predicting variance. PCs are augmented online by perturbing the points with Gaussian noise and random rotations ($\pm 15^\circ$). Finally, the PCs are median-centred to origin and normalised to have the same surface area. The networks are trained until convergence using an Adam optimiser with an initial learning rate of 5×10^{-4} . Specific to the VAE, we use KL-annealing by increasing β from 0 to 0.1.

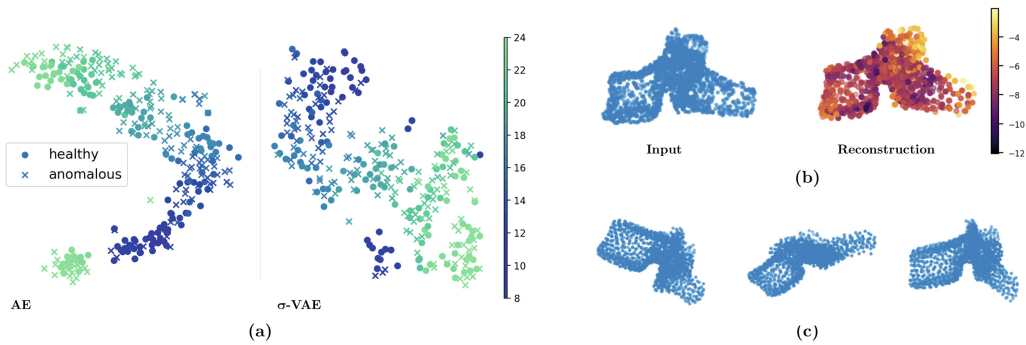


Fig. 4. Characteristics of σ -VAE: (a) Comparison of TSNE embeddings of simple p AE with σ -VAE. Observe transition in clusters being inline with vertebral indices. Note that embedding becomes compact for a VAE. (b) A PC and its reconstruction coloured with $\log(\sigma^2)$ of every point. Observe high variance in vertebral processes. (c) Example generations from decoder with $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. (Color figure online)

Qualitative Evaluation of AE Architectures. We investigate if meaningful shape features can be learnt without supervision. Validating this, in Fig. 4a, we plot a TSNE embedding of the test set latent vectors learnt by a naive p AE and σ -VAE trained only on healthy vertebrae. Observe the clusters formed based on the vertebral index and the transition between the indices. This corresponds to the natural variation of vertebral shapes in a human spine. Indicating the fractured vertebrae in the embedding, we highlight their degree of similarity with the healthy counterparts. Also, observe that embedding is more regularised representing a Gaussian in case of σ -VAE, indicating the continuity of the learnt latent space. Figure 4b shows the predictive variance modelled by the σ -VAE. Posterior elements of a vertebrae are the most varying among population. Observe this being captured by the variance in the vertebral process regions. Lastly, illustrating σ -VAE's generative capabilities, Fig. 4c shows vertebral PC samples generated by sampling the latent vector, $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$.

Vertebral Fracture Detection. Evaluating the reconstruction quality of our p AE architectures, we employ them to detect fractures as anomalies. As baselines, we choose two supervised approaches: (1) point-net (PN), the encoding

part in our p AE architectures, cast as a binary classifier and (2) the same point-net trained with median frequency balancing the classes (ref. as PN_{bal}) to accentuate the loss from minority fractured class. We report their performance in Table 1, over 3-fold cross-validation while retaining the ratio of healthy to fractured vertebrae in the data splits. Frequency balancing improves the F1 score significantly, albeit not at the level of the proposed anomaly detection schemes.

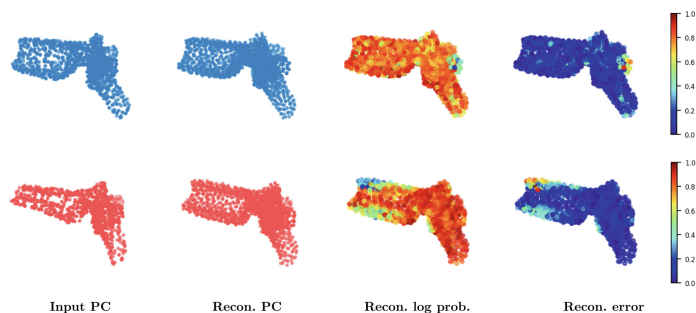


Fig. 5. Reconstructions: healthy (top) and fractured (bottom) vertebral PCs. Observe p AE’s ‘healthy’ reconstruction of a fracture. Errors and log-probabilities are normalised to $[0, 1]$ within PC for visualisation, but anomaly detection works on un-normalised values.

Table 1. Performance comparison of unsupervised and supervised fracture detection approaches. Measures: Precision (P), Recall (R), F1-score, and area-under-ROC curve (AUC) computed by varying thresholds on recon. error and recon. log-probabilities. Since supervised models have no threshold selection, AUC is not reported.

Measures	PN	PN_{bal}	<i>recon. error</i>				<i>recon. log-likelihood</i>	
			AE	VAE	σ -AE	σ -VAE	σ -AE	σ -VAE
P	100 ± 0.0	68.6 ± 3.4	57.6 ± 4.1	61.1 ± 1.9	67.1 ± 6.5	68.4 ± 3.3	62.3 ± 4.3	61.6 ± 1.4
R	13.9 ± 3.1	57.6 ± 7.5	85.0 ± 9.8	79.0 ± 3.6	74.3 ± 4.0	71.7 ± 4.1	72.7 ± 6.1	79.7 ± 2.5
$F1$	24.7 ± 4.7	62.5 ± 5.8	68.0 ± 0.9	68.5 ± 1.7	67.5 ± 5.1	69.6 ± 1.2	66.7 ± 1.3	69.5 ± 0.6
AUC	n.a	n.a	70.8 ± 2.2	74.8 ± 3.0	75.9 ± 2.0	75.9 ± 1.5	70.2 ± 2.2	73.8 ± 2.0

Reconstruction for fracture detection: When detecting fractures based on reconstruction error (d_{ch}), we observe that a naive p AE already out-performs the supervised classifiers (cf. Table 1). On top of this, we see that latent space modelling and variance modelling individually offer an improvement in F1-scores while increasing the AUC, indicating a stable detection of fractures. The performance of both σ -AE and σ -VAE is similar indicating the role of loss attenuation. However, the advantage of explicitly regularising the latent space for σ -VAE can be seen in likelihood-based anomaly detection, where σ -VAE outperforms σ -AE. Figure 5 compares a reconstruction of a healthy and fractured vertebrae of the same vertebral level. Note the high reconstruction error and a low log-likelihood spatially corresponding to the deformity due to fracture.

4 Conclusions

We presented point-cloud-based auto-encoding architectures for extracting descriptive shape features. Improving their description, we incorporated variance and latent space-modelling capability using specially defined PC specific losses. The former captures the natural variance in the data while the latter regularises the latent space to be continuous. Deploying these networks for the task of unsupervised fracture detection, we achieved an AUC of 76% without using any intensity or textural features. Future work will combine the extracted shape signatures with textural features e.g. bone density and trabecular texture of vertebrae to perform fracture-grade classification.

Acknowledgements. 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 Quadro P5000 used for this work was donated by NVIDIA Corporation.

References

1. Ingalhalikar, M., et al.: Sex differences in the structural connectome of the human brain. *Proc. Natl. Acad. Sci.* **111**(2), 823–828 (2014)
2. Shakeri, M., Lombaert, H., Tripathi, S., Kadoury, S.: Deep spectral-based shape features for Alzheimer’s disease classification. In: Reuter, M., Wachinger, C., Lombaert, H. (eds.) *SeSAMI 2016*. LNCS, vol. 10126, pp. 15–24. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-51237-2_2
3. Isensee, F., et al.: Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge. *arXiv e-prints* (2018)
4. Bronstein, M.M., et al.: Geometric deep learning: going beyond euclidean data. *IEEE Sig. Process. Mag.* **34**(4), 18–42 (2017)
5. Qi, C.R., et al.: PointNet: deep learning on point sets for 3D classification and segmentation. In: *CVPR* (2017)
6. Gutiérrez-Becker, B., Wachinger, C.: Deep multi-structural shape analysis: application to neuroanatomy. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *MICCAI 2018*. LNCS, vol. 11072, pp. 523–531. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_60
7. Fan, H., et al.: A point set generation network for 3D object reconstruction from a single image. In: *CVPR* (2017)
8. Yang, Y., et al.: FoldingNet: point cloud auto-encoder via deep grid deformation. In: *CVPR* (2018)
9. Baum, T., et al.: Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT. *Eur. Radiol.* **24**(4), 872–880 (2014)
10. Tomita, N., et al.: Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput. Biol. Med.* **98**, 8–15 (2018)
11. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: *NIPS* (2017)
12. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Technical report, SNU Data Mining Center (2015)

PART III

CONCLUDING REMARKS



Discussion

The central theme of this thesis is the advancement of automated spine image analysis with a special focus on learning and incorporating structural priors from the data. In the spirit of a cumulative thesis, these contributions are presented in Chapters 5–8, in the form of self-contained, peer-reviewed conference or journal articles. Specifically, Chapters 5, 6, and 7 tackle with the problem of *vertebrae labelling* and Chapter 8 tackles the problem of *vertebral fracture detection*. As each of these chapters also *discussed and concludes* its corresponding contribution, this section will be *discuss* their key contributions in the broader context of spine image analysis while detailing the questions that these contributions opened up.

Chapter 5 - Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior

Accurate localisation and identification of vertebrae is a challenging problem in spine image analysis, especially due to the need to combine global context with localised feature extraction. However, a larger receptive field implied the CNN to be deeper, thereby making it heavy, more so, when processing three-dimensional data. The proposed combination of Btrfly Net and an energy-based discriminator validated the use of 2D orthogonal projections for accurate vertebrae labelling. Especially, the role of adversarial-training in learning a local anatomical prior was validated quantitatively and qualitatively.

Although Btrfly Net is tailored towards the spine due to it working with maximum intensity projections, the composite of a CNN as a generator (G) and an auto-encoder as an energy-based discriminator (D) generalises to other applications. However, since the autoencoder is fully convolutional with a limited receptive field, it learns a local anatomical prior, which is not always useful. E.g. in cephalometric x-rays, a global prior is more valuable than a local prior.

Chapter 6 - Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy

Investigating how a local anatomical prior is enforced into the Btrfly Net in an energy-based adversarial, the representations learnt by the Btrfly Net thus trained are analysed. Furthermore, the possibility of adversarially enforcing a global spine prior is studied, thanks to a discriminator with a densely-connected last layer. The dense layer increases the receptive field to the full input. Additionally, addressing the problem of occluded projections in clinically occurring scans, a spine localisation network is employed to obtain improved localised projections. It is concluded that the proposed energy-based prior learning, whose receptive field is fixed, enforces structure over a local region, irrespective of the scan's FoV. On the other hand, learning a global prior is challenging due to high variability in the FoV, thus rendering its enforcement not as effective. In spite of its limited success in spine, it must be noted that global prior learning has potential in other data domains such as cephalometric x-rays or hand x-rays.

Chapter 7 - Pushing the limits of an FCN and a CRF towards near-ideal vertebrae labelling

Observe that in the previous two chapters, the learnt priors are adversarially enforced *ad-hoc* into the primary CNN, the Btrfly Net. In this chapter, the limits of traditional, *post-hoc* prior enforcement approaches were investigated. An important reason for this was to accommodate the naturally occurring anatomical abnormalities in the spine (Sec. 2.2), which are challenging to learn in a data-driven manner due to their rare occurrence. Additionally, the 2D Btrfly Net was replaced with a more generalisable 3D FCN architecture. This FCN-CRF combination achieved high performance on multiple public datasets, in a fully automated setting, including the selection of the correct prior model for the spine from among five anatomical variations. The linear CRF is capable of updating its prediction by solving a linear system in real-time, thus resulting in fast runtime. Thus, an interaction module was developed to take minimal expert-input (e.g. choose the right prior model, adjust one wrong prediction etc.) to correct erroneous predictions in real time. This human-machine combination achieved near-ideal labelling performance.

Chapter 8 - Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis

Switching from the processing end to the diagnosing end, the problem of fracture detection is tackled from a shape modelling perspective, while addressing the challenges of data representation and limited annotations. First, the vertebral shapes are represented as point clouds, a very efficient representation compared to voxels. Compared to other surface presentations such as meshes, point clouds are easier to work with, thanks to their permutation-invariant representation. Then, the problem of fracture detection is posed as a problem of outlier detection, wherein the distribution of healthy vertebrae is learnt using a VAE and a fracture is detected as an outlier in this distribution. Furthermore, the VAE was extended to learn the data variance at a point-level in the reconstruction space, which enabled localisation of the outlying region, making the proposed approach interpretable.

Due to the domain agnostic nature of the data representation, the proposed approach generalises across modalities and across morphology. An important limitation of the proposed approach is that the distribution is learnt unconditionally, wherein, the distribution of all the vertebrae, be it thoracic or lumbar, are mapped to a unimodal Gaussian. This is restrictive and a multi-modal mapping conditioned on the vertebrae label could result in a superior performance.



Outlook

Deep neural networks are immensely capable of learning task-effective representations given *sufficient* data. However, the sufficiency of data is ill-defined. Moreover, medical images not only lack the semantic information typical to natural images, but contain a heavy inductive bias. How many annotated cases does a model need to learn reliable representations? This brings in unique challenges towards clinical deployment where domain generalisation, calibrated confidence prediction, and explainability are more important. A wrong prediction is just more expensive. This leads us towards learning data priors and explicitly or implicitly incorporating them into the learning systems. In this thesis' attempt towards this, several novel research problems opened up and progress occurred in other fields that could act as an inspiration for future research in spine image analysis. Again as with the rest of the thesis, we categorise the outlook into the two parts of vertebrae labelling and fracture detection, both using data priors.

Vertebrae labelling

Taking the baton from CNNs, transformer-based architectures are currently the chosen approach for image processing. SpineTransformers, proposed by Tao et al. [5], are their first successful application in spine image analysis. The attention-layer of the transformer acts as a proxy for learning and incorporating an anatomical prior, but this is not explicit. An CNN-based implemented with explicit prior incorporation eventually showed superior performance [30]. An important research direction could thus be the study of prior incorporation in transformer-based architecture.

Taking a step back, we look at prior enforcement itself. In every approach till date, the prior has been learnt from data, be it using autoencoders or conditional random fields. However, the errors we face in vertebrae labelling are more simpler, e.g. L1 detected above T12. Or L1 and T12 labelled on the same vertebrae. These simple errors can be sorted using hard-coded rules. Alternatively, these rules can be posed as constraints in a combinatorial optimisation problem [45], whose loss can be

used as a regulariser to train the labelling network. This domain of combinatorial prior enforcement is another interesting research venue.

Lastly, a problem that is yet to be solved in literature is the accurate labelling of vertebrae in the presence of transitional vertebrae (T13 or L6) or in spines with other naturally occurring anomalies. The strong inductive prior of a normal spine learnt by a data-driven model and the rarity of abnormal spines (with transitional vertebrae) makes this problem a difficult one to solve. Further research in shape analysis, uncertainty quantification, and prior enforcement is needed to address this problem.

Fracture detection

Admittedly, generative modelling has seen immense progress in recent years, e.g improved GANs [46] and VAEs [47], normalising flows [15], and diffusion models [16]. Almost all these models have also percolated into the point cloud domain [48, 49, 50]. An obvious extension of [51] would be to model the vertebral shapes, conditionally, using improved generative models.

Furthermore, outlier detection, or its parent field, generalised out-of-distribution (OOD) detection [52], offers multiple avenues for improved fracture detection. It is also important to note that the fractures are not the only outliers from a normal distribution. Vertebrae with screws, Schmorl’s node etc. also count as outliers and cannot be distinguished under the proposed approach. Outlier detection in this scenario is called multi-class anomaly detection [53], an interesting research direction. The problem of fracture grading makes this problem all the more interesting.

Spine image analysis, in general

Through this thesis, we have only scratched a surface of automated spine image analysis. On the image-processing end, the presented labelling approaches can be improved, adapted, and translated to other imaging modalities such as MR [54] and 2D radiographs [55]. These modalities also bring research problems of their own [56, 57]. Efficient 3D image processing is always of interest, given the size of a spine scan using resource efficient neural networks such as Binarised CNNs [58] or quantised CNNs [59]. On a clinical front, automated spine analysis has potential for extracting crucial biomarkers with global consequences, for instance, towards osteoporosis detection [60], spinal metastasis [61], back pain [62] etc.

All in all, the field of automated spine image analysis is all the more active today, with challenging research questions both on clinical and methodological fronts, and we hope this thesis plays its small part in advancing the field forward.



Bibliography

- [1] E. von der Lippe, L. Krause, M. Porst, A. Wengler, J. Leddin, A. Müller, M.-L. Zeisler, A. Anton, A. Rommel, B. 2. study group, et al. “Prevalence of back and neck pain in Germany. Results from the BURDEN 2020 Burden of Disease Study.” In: *Journal of Health Monitoring* 6.Suppl 3 (2021), p. 2.
- [2] K. T. Palmer, K. Walsh, H. Bendall, C. Cooper, and D. Coggon. “Back pain in Britain: comparison of two prevalence surveys at an interval of 10 years.” In: *Bmj* 320.7249 (2000), pp. 1577–1578.
- [3] J. Cauley, D. Thompson, K. Ensrud, J. Scott, and D. Black. “Risk of mortality following clinical fractures.” In: *Osteoporosis international* 11 (2000), pp. 556–561.
- [4] A. Sekuboyina, M. E. Husseini, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern, et al. “VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images.” In: *Medical image analysis* 73 (2021), p. 102166.
- [5] R. Tao, W. Liu, and G. Zheng. “Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers.” In: *Medical Image Analysis* 75 (2022), p. 102258.
- [6] M. Husseini*, A. Sekuboyina*, M. Loeffler, F. Navarro, B. H. Menze, and J. S. Kirschke. “Grading loss: a fracture grade-based metric loss for vertebral fracture detection.” In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer. 2020, pp. 733–742.
- [7] W. Brinjikji, P. H. Luetmer, B. Comstock, B. W. Bresnahan, L. Chen, R. Deyo, S. Halabi, J. Turner, A. Avins, K. James, et al. “Systematic literature review of imaging features of spinal degeneration in asymptomatic populations.” In: *American journal of neuroradiology* 36.4 (2015), pp. 811–816.

- [8] A. Sekuboyina, M. Rempfler, J. Kukačka, G. Tetteh, A. Valentinič, J. S. Kirschke, and B. H. Menze. “Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior.” In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*. Springer. 2018, pp. 649–657.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks.” In: *Communications of the ACM* 60.6 (2017), pp. 84–90.
- [10] F. Rosenblatt. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Tech. rep. Cornell Aeronautical Lab Inc Buffalo NY, 1961.
- [11] Y. LeCun, Y. Bengio, et al. “Convolutional networks for images, speech, and time series.” In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.
- [12] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation.” In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.
- [13] J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.
- [14] F. Milletari, N. Navab, and S.-A. Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation.” In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.
- [15] D. Rezende and S. Mohamed. “Variational inference with normalizing flows.” In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.
- [16] Y. Song, C. Durkan, I. Murray, and S. Ermon. “Maximum likelihood training of score-based diffusion models.” In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1415–1428.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial networks.” In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

-
- [18] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein generative adversarial networks.” In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.
- [19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. “Improved training of wasserstein gans.” In: *Advances in neural information processing systems* 30 (2017).
- [20] J. Zhao, M. Mathieu, and Y. LeCun. “Energy-based generative adversarial network.” In: *arXiv preprint arXiv:1609.03126* (2016).
- [21] D. P. Kingma and M. Welling. “Auto-encoding variational bayes.” In: *arXiv preprint arXiv:1312.6114* (2013).
- [22] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. “beta-vae: Learning basic visual concepts with a constrained variational framework.” In: *International conference on learning representations*. 2017.
- [23] A. Bayat, A. Sekuboyina, F. Hofmann, M. E. Hussein, J. S. Kirschke, and B. H. Menze. “Vertebral labelling in radiographs: learning a coordinate corrector to enforce spinal shape.” In: *Computational Methods and Clinical Applications for Spine Imaging: 6th International Workshop and Challenge, CSI 2019, Shenzhen, China, October 17, 2019, Proceedings 6*. Springer. 2020, pp. 39–46.
- [24] Y. Sun, X. Wang, and X. Tang. “Deep convolutional network cascade for facial point detection.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3476–3483.
- [25] A. Toshev and C. Szegedy. “Deeppose: Human pose estimation via deep neural networks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1653–1660.
- [26] W. Li, Y. Lu, K. Zheng, H. Liao, C. Lin, J. Luo, C.-T. Cheng, J. Xiao, L. Lu, C.-F. Kuo, et al. “Structured landmark detection via topology-adapting deep graph learning.” In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 266–283.
- [27] A. Newell, K. Yang, and J. Deng. “Stacked hourglass networks for human pose estimation.” In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 483–499.

- [28] K. Sun, B. Xiao, D. Liu, and J. Wang. “Deep high-resolution representation learning for human pose estimation.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5693–5703.
- [29] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. “Deep high-resolution representation learning for visual recognition.” In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), pp. 3349–3364.
- [30] A. Sekuboyina*, J. Irmair*, S. Shit, J. Kirschke, B. Andres, and B. Menze. “Pushing the limits of an FCN and a CRF towards near-ideal vertebrae labelling.” In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5.
- [31] 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: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part III 15*. Springer. 2012, pp. 590–598.
- [32] H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J. C. Cheng, and P.-A. Heng. “Automatic localization and identification of vertebrae in spine CT via a joint learning model with deep neural networks.” In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer. 2015, pp. 515–522.
- [33] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu. “Test-time adaptable neural networks for robust medical image segmentation.” In: *Medical Image Analysis* 68 (2021), p. 101907.
- [34] A. J. Larrazabal, C. Martínez, B. Glocker, and E. Ferrante. “Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders.” In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 3813–3820.
- [35] A. J. Larrazabal, C. Martinez, and E. Ferrante. “Anatomical priors for image segmentation via post-processing with denoising autoencoders.” In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer. 2019, pp. 585–593.

-
- [36] S. Shit*, J. C. Paetzold*, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylyka, J. P. Pluim, U. Bauer, and B. H. Menze. “clDice-a Novel Topology-Preserving Loss Function for Tubular Structure Segmentation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16560–16569.
- [37] 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.” In: *IEEE transactions on medical imaging* 37.2 (2017), pp. 384–395.
- [38] B. Gutiérrez-Becker and C. Wachinger. “Deep multi-structural shape analysis: application to neuroanatomy.” In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part III 11*. Springer. 2018, pp. 523–531.
- [39] J. Yu, C. Zhang, H. Wang, D. Zhang, Y. Song, T. Xiang, D. Liu, and W. Cai. “3d medical point transformer: Introducing convolution to attention networks for medical point cloud analysis.” In: *arXiv preprint arXiv:2112.04863* (2021).
- [40] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. “Pointnet: Deep learning on point sets for 3d classification and segmentation.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.
- [41] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. “Pointnet++: Deep hierarchical feature learning on point sets in a metric space.” In: *Advances in neural information processing systems* 30 (2017).
- [42] J. An and S. Cho. “Variational autoencoder based anomaly detection using reconstruction probability.” In: *Special lecture on IE* 2.1 (2015), pp. 1–18.
- [43] R. Yao, C. Liu, L. Zhang, and P. Peng. “Unsupervised anomaly detection using variational auto-encoder based feature extraction.” In: *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*. IEEE. 2019, pp. 1–7.
- [44] A. Vasilev, V. Golkov, M. Meissner, I. Lipp, E. Sgarlata, V. Tomassini, D. K. Jones, and D. Cremers. “q-Space novelty detection with variational autoencoders.” In: *Computational Diffusion MRI: MICCAI Workshop, Shenzhen, China, October 2019*. Springer. 2020, pp. 113–124.

- [45] M. V. Pogančić, A. Paulus, V. Musil, G. Martius, and M. Rolinek. “Differentiation of blackbox combinatorial solvers.” In: *International Conference on Learning Representations*. 2020.
- [46] X. Yu, X. Zhang, Y. Cao, and M. Xia. “VAEGAN: A Collaborative Filtering Framework based on Adversarial Variational Autoencoders.” In: *IJCAI*. 2019, pp. 4206–4212.
- [47] A. Van Den Oord, O. Vinyals, et al. “Neural discrete representation learning.” In: *Advances in neural information processing systems* 30 (2017).
- [48] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. “Point-flow: 3d point cloud generation with continuous normalizing flows.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4541–4550.
- [49] Z. Han, X. Wang, Y.-S. Liu, and M. Zwicker. “Multi-angle point cloud-VAE: Unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction.” In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. 2019, pp. 10441–10450.
- [50] S. Luo and W. Hu. “Diffusion probabilistic models for 3d point cloud generation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2837–2845.
- [51] A. Sekuboyina, M. Rempfler, A. Valentinitzsch, M. Loeffler, J. S. Kirschke, and B. H. Menze. “Probabilistic point cloud reconstructions for vertebral shape analysis.” In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer. 2019, pp. 375–383.
- [52] J. Yang, K. Zhou, Y. Li, and Z. Liu. “Generalized out-of-distribution detection: A survey.” In: *arXiv preprint arXiv:2110.11334* (2021).
- [53] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le. “A Unified Model for Multi-class Anomaly Detection.” In: *arXiv preprint arXiv:2206.03687* (2022).
- [54] R. Windsor, A. Jamaludin, T. Kadir, and A. Zisserman. “A convolutional approach to vertebrae detection and labelling in whole spine MRI.” In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer. 2020, pp. 712–722.

-
- [55] A. Bayat, D. F. Pace, A. Sekuboyina, C. Payer, D. Stern, M. Urschler, J. S. Kirschke, and B. H. Menze. “Anatomy-aware inference of the 3D standing spine posture from 2D radiographs.” In: *Tomography* 8.1 (2022), pp. 479–496.
- [56] J. Dolz, C. Desrosiers, and I. Ben Ayed. “IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet.” In: *Computational Methods and Clinical Applications for Spine Imaging: 5th International Workshop and Challenge, CSI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*. Springer. 2019, pp. 130–143.
- [57] 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: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer. 2020, pp. 775–784.
- [58] X. Lin, C. Zhao, and W. Pan. “Towards accurate binary convolutional neural network.” In: *Advances in neural information processing systems* 30 (2017).
- [59] Z. Cai, X. He, J. Sun, and N. Vasconcelos. “Deep learning with low precision by half-wave gaussian quantization.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5918–5926.
- [60] S. Kim, B. R. Kim, H.-D. Chae, J. Lee, S.-J. Ye, D. H. Kim, S. H. Hong, J.-Y. Choi, and H. J. Yoo. “Deep Radiomics-based Approach to the Diagnosis of Osteoporosis Using Hip Radiographs.” In: *Radiology: Artificial Intelligence* 4.4 (2022), e210212.
- [61] W. Ong, L. Zhu, W. Zhang, T. Kuah, D. S. W. Lim, X. Z. Low, Y. L. Thian, E. C. Teo, J. H. Tan, N. Kumar, et al. “Application of Artificial Intelligence Methods for Imaging of Spinal Metastasis.” In: *Cancers* 14.16 (2022), p. 4025.
- [62] S. D. Tagliaferri, M. Angelova, X. Zhao, P. J. Owen, C. T. Miller, T. Wilkin, and D. L. Belavy. “Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews.” In: *NPJ digital medicine* 3.1 (2020), p. 93.

PART IV

APPENDICES



**Supplementary Material: Btrfly
Net: Vertebrae Labelling with
Energy-Based Adversarial
Learning of Local Spine Prior**

5 Supplementary Material

5.1 Case study on a non-spine-centred scan

The benchmark dataset used in Section 3 of our work is mostly spine-centred, and the naive maximum intensity projections contain no occlusions. However, in certain full-body scans, the spine is obstructed by the ribcage in a MIP of the entire scan, or the spine is not spatially centred in both the views, thus not taking full advantage of Btrfly net’s view fusion (cf. Fig. 6a). Such cases can be handled by introducing a pre-processing step before the Btrfly_{pe} net in the form of an ‘object-detection’ network.

For such scenario, we construct the MIPs in two stages. The first MIP is constructed on the entire scan. On this, we use a *single-shot object detection* (SSD) inspired architecture [1] trained to identify occluded spines (cf. Fig. 6a). Once the spine is located, we construct the second pair of MIPs based on the *spine-slices*, which are then used as inputs to the Btrfly_{pe} net (cf. Fig. 6b,c). The ground truth for the SSD net can be constructed from the ground truth annotation of the vertebral centroids. We use a generic 16-layer residual CNN with an SSD extension. This use-case is illustrated on a scan from the training set of the xVertSeg [2] dataset. Note that we used the xVertSeg data only for inference and not for re-training the network. The centroids of the vertebrae are obtained from the maximum point of the distance transform of the segmentation map (xVertSeg has voxel-level annotations from L1 to L5).

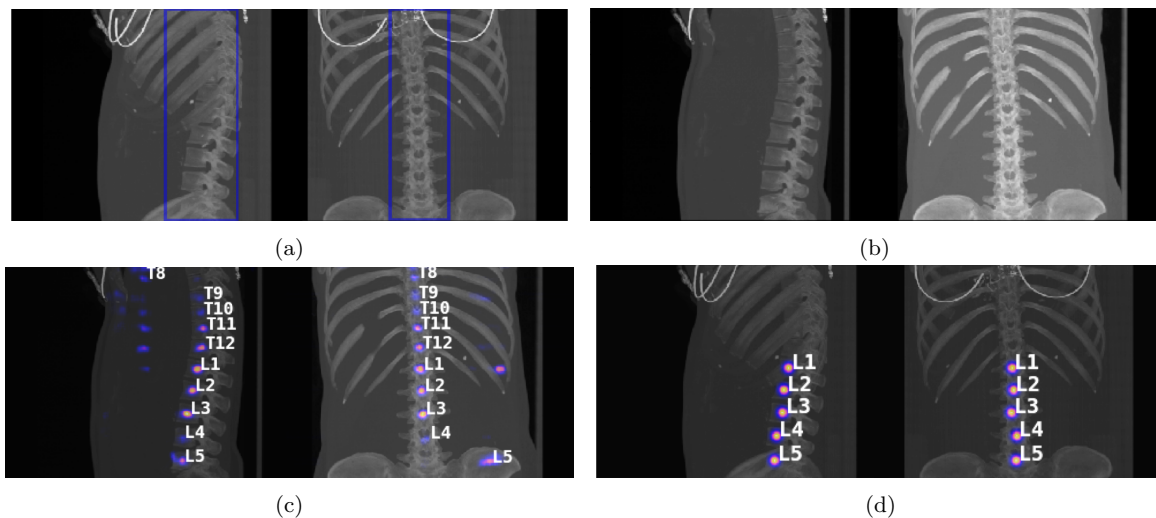


Fig. 6: An illustration of the extension to Btrfly_{pe} net. (a) Naive sagittal and coronal MIPs on the entire scan with the bounding box predictions (in blue) of our SSD net. Observe the ribcage obstructing the spine. (b) Improved MIPs constructed from the slices containing the spine based on the localisation in (a). (c) Output of the Btrfly_{pe} net, resulting in an 80 % id.rate. Also observe the incorrect localisation of T8 and L5, along with prediction noise in sagittal view owing to the non-aligned spine in both views. We believe that aligning the spine using its detection could further improve the prediction. (d) The ground truth centroids constructed from the voxel-level annotation map of scan. Since xVertSeg data only has lumbar annotations, we visualise the lumbar centroids.

References

1. Liu W. et al.: SSD: Single Shot MultiBox Detector. CoRR abs/1512.02325 (2015), <http://arxiv.org/abs/1512.02325>.
2. The xVertSeg Challenge, <http://lit.fe.uni-lj.si/xVertSeg/>

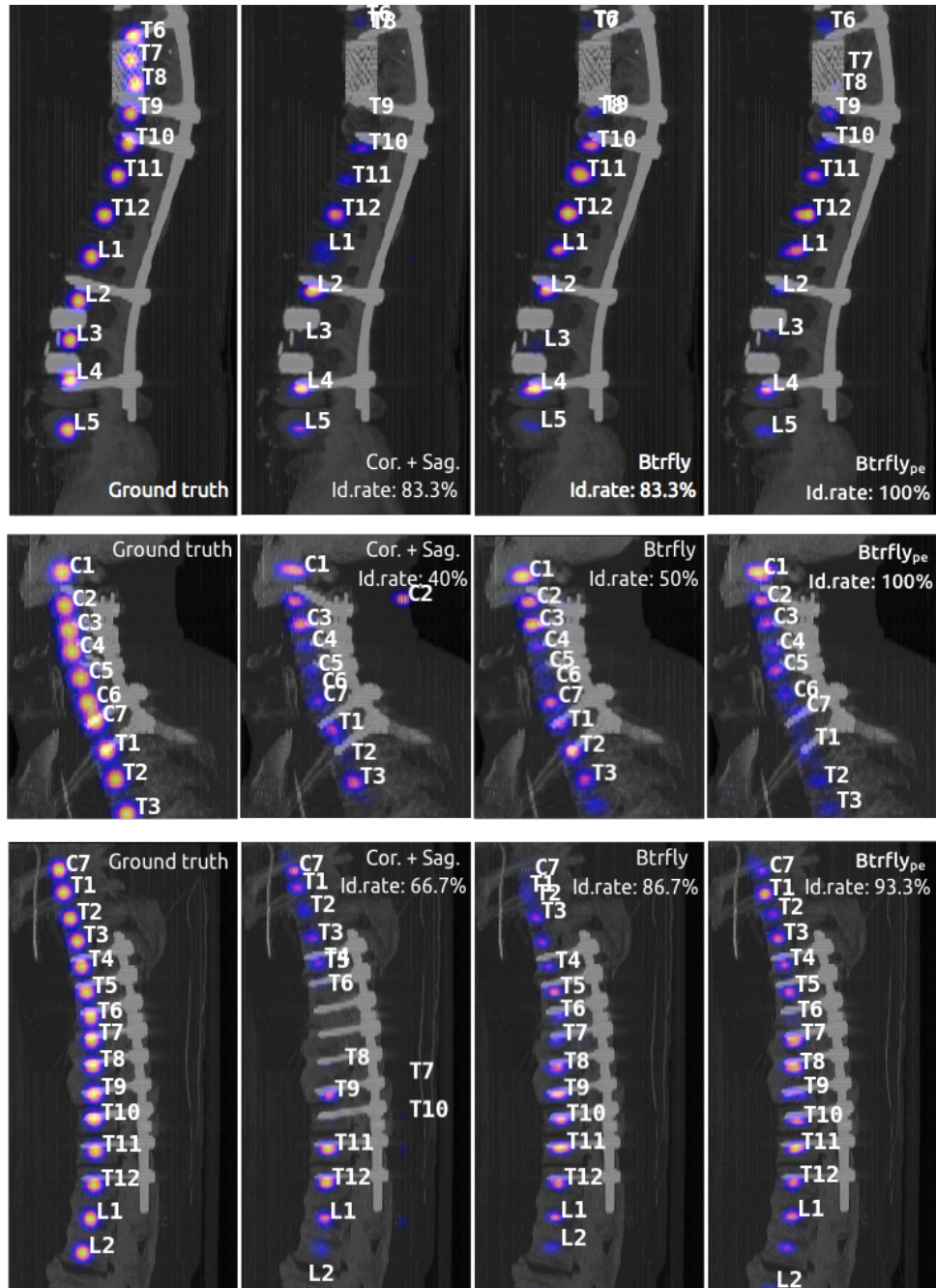


Fig. 7: **Additional quantitative results.** MIP images with predictions of the three variants of our approach at $T=0$ for all cases. The spine's local structure is conserved in the predictions of Btrfly_{pe}. Also observe that, as a consequence of prior encoding, in some cases labels are predicted in spite of no useful spatial information, albeit the strength of these predictions is less.



**Supplementary Material: Labeling
Vertebrae with Two-dimensional
Reformations of Multidetector CT
Images: An Adversarial Approach
for Incorporating Prior Knowledge
of Spine Anatomy**

Appendix E1

In this section, we present the architectural and training details of various components in our approach in three parts: spine localization, Btrfly network, and prior encoding using adversarial learning.

Spine Localization

For localization, annotations were obtained from the available vertebral centroids by putting Gaussians at vertebral locations. The input was a 3D scan at low isotropic resolution of 4 mm. The output was also a 3D volume with a Gaussian at every vertebra location. We used a higher σ ($= 15$) for wider Gaussians as annotations. The architecture was a light-weight U-Net as shown in Figure E1. The loss used for training is a simple ℓ_2 regression loss.

Training

For training, an Adam optimizer was employed with an initial learning rate of 5×10^{-4} . Convergence was tested on a held-out validation set consisting of 10% of the training data. Once trained, a bounding box is constructed out of the predicted heatmap using a fixed tolerance around the “active” voxels in the prediction, which indicate the presence of the spine.

Btrfly Net

The detailed architecture of the Btrfly net is shown in Figure E2. Similar to that in Sekuboyina et al (11), the loss for the Btrfly net is a combination of cross-entropy and ℓ_2 regression loss.

Denoting \mathbf{x}_{view} and \mathbf{y}_{view} , $view \in \{\text{sag}, \text{cor}\}$ to be the 2D projections of the image and the annotations respectively, the loss of the sagittal arm of the Btrfly Net was given by:

$$\mathcal{L}_{b,sag} = \|\mathbf{y}_{sag} - \tilde{\mathbf{y}}_{sag}\|_2 + \omega H(\mathbf{y}_{sag}^\sigma, \tilde{\mathbf{y}}_{sag}^\sigma),$$

where $\tilde{\mathbf{y}}_{sag}$ is the prediction of the net's xz-arm, H is the cross-entropy function, and $\mathbf{y}_{sag}^\sigma = \sigma(\mathbf{y}_{sag})$, the softmax excitation of the prediction. ω is the median frequency weight map over the occurrence of vertebral labels, incorporated for boosting the learning of less frequent vertebral classes. The loss for the yz-arm was similarly constructed, resulting in the total loss of the Btrfly net:

$$\mathcal{L}_{btrfly} = \mathcal{L}_{b,sag} + \mathcal{L}_{b,cor}.$$

Training

Data for this phase consisted of MIPs across the two views. For training the network, an Adam optimizer was employed with an initial learning rate of $\lambda = 1 \times 10^{-3}$. λ was decayed by a factor of 3/4th every 10k iterations till 0.2×10^{-3} . 10% of the training data were held out to ascertain training convergence.

Prior Learning Using Adversaries

In this article, we also investigated the ability of two adversaries in aiding the labeling of vertebrae: an energy-based adversary and a Wasserstein-distance based adversary. Both the adversaries worked on the label maps (\mathbf{y}_{view} or $\tilde{\mathbf{y}}_{view}$) and provided a discriminating signal specific to their architecture. This signal corresponded to “real-ness” or “fakeness” of the input to the discriminator. Figure E3 shows the arrangement of the discriminators with respect to the Btrfly Net and their detailed architecture.

EB-D

The discriminating signal here was the ℓ_2 distance between the input \mathbf{y}_{view} (or $\tilde{\mathbf{y}}_{view}$) and its reconstruction, $rec(\mathbf{y}_{view})$: $\mathcal{D}(\mathbf{y}_{view}) = E = \|\mathbf{y}_{view} - rec(\mathbf{y}_{view})\|_2$. The total loss combined with the Btrfly net (generator, G) is as follows:

$$\mathcal{L}_D = \mathcal{D}(\mathbf{y}_{view}) + \max(0, m - \mathcal{D}(\tilde{\mathbf{y}}_{view})), \text{ and}$$
$$\mathcal{L}_G = \mathcal{D}(\tilde{\mathbf{y}}_{view}) + \mathcal{L}_{b,view}.$$

Thus, D learned to better reconstruct “real” label maps by minimizing the reconstruction loss for a real sample. On the other hand, G learned to predict better “fake” label maps so that the discriminator gives a low reconstruction loss. m in the equation above is a margin discouraging EB-D from learning to reconstruct fake or predicted samples. It was varied from 10 to 0 rapidly (m was halved every 1000 iterations).

W-D

In the case of Wasserstein GAN, the D- or G-specific losses are not as intuitive as in the earlier case. However, the combined setup of the generator (Btrfly) and discriminator tried to minimize the Wasserstein distance between the distributions of real and generated label maps. We followed the loss design as in the improved Wasserstein GAN with gradient penalty (17) as below:

$$\mathcal{L}_D = \mathcal{D}(\tilde{\mathbf{y}}_{view}) - \mathcal{D}(\mathbf{y}_{view}) + \lambda(\|\nabla_{\hat{\mathbf{y}}_{view}} \mathcal{D}(\mathbf{y}_{view})\|_2 - 1)^2, \text{ and}$$
$$\mathcal{L}_G = -\mathcal{D}(\tilde{\mathbf{y}}_{view}) + \mathcal{L}_{b,view},$$

where $\hat{\mathbf{y}}_{view} = \epsilon \mathbf{y}_{view} + (1 - \epsilon) \tilde{\mathbf{y}}_{view}$, for any $\epsilon \sim U[0,1]$. λ (= 1.0 in our work) is a scalar weight for the regularization term on the gradients.

Reference

17. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. Poster presented at: Advances in Neural Information Processing Systems 30; December 4-9, 2017, Long Beach, CA. <https://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans>. Accessed March 9, 2020.

Supplemental Figures

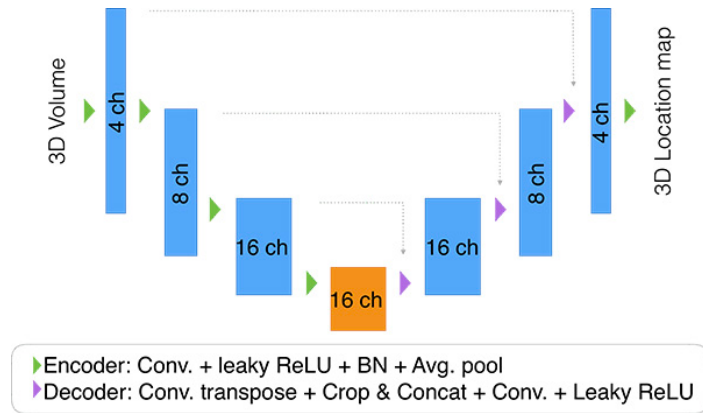


Figure E1: Localization network architecture. Kernel sizes for all convolution kernels except the last layer are $3 \times 3 \times 3$. Last convolutional kernel is $1 \times 1 \times 1$. Transposed convolution kernels are $4 \times 4 \times 4$. Average pooling kernels are $2 \times 2 \times 2$ with a stride of 2.

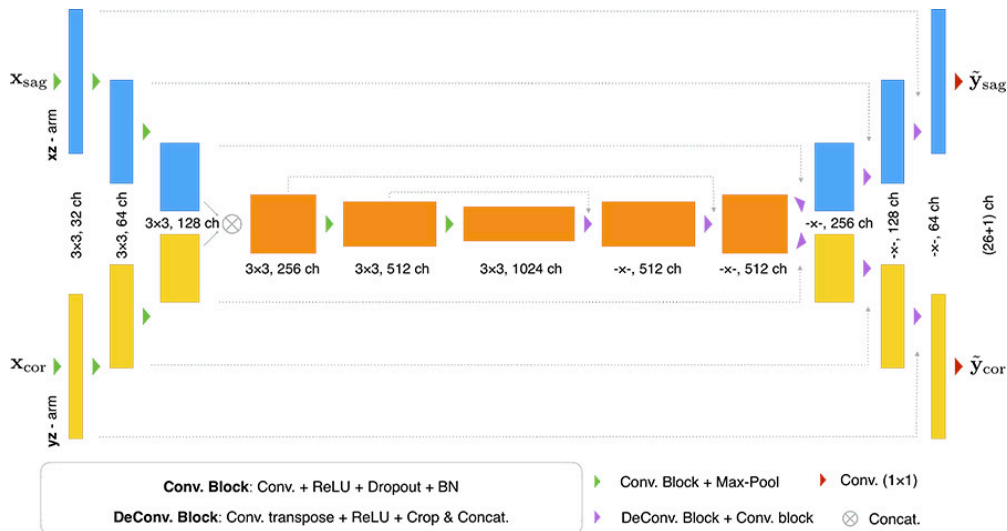


Figure E2: Btrfly-Net architecture. The xz- (blue) and the yz arms (yellow) correspond to the sagittal and coronal views. In upscaling path, kernel size \times denotes two different kernel sizes: 4×4 transposed convolution followed by 3×3 convolution.

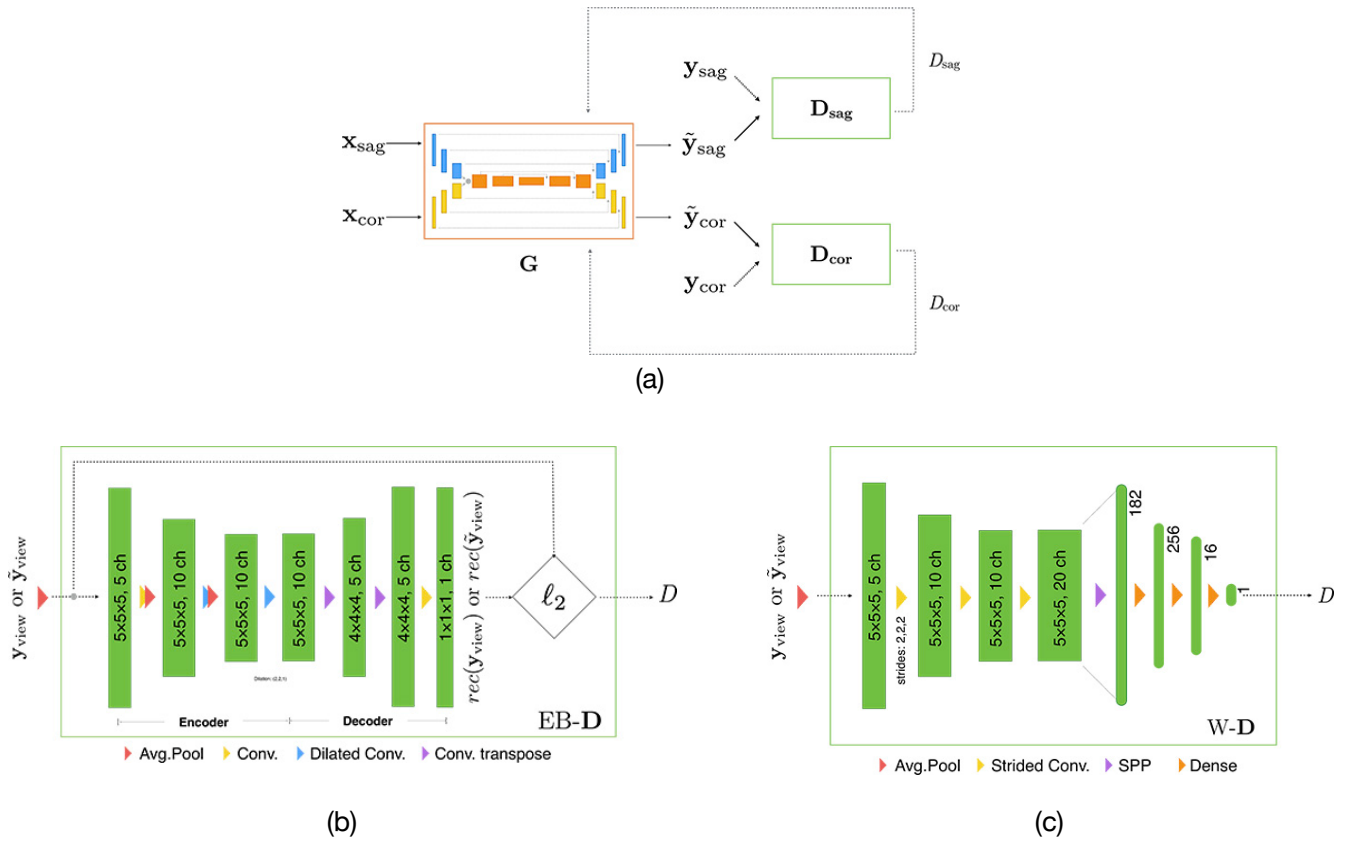


Figure E3: (a) The arrangement of the discriminators with respect to the Butterfly net. (b) The architecture of the Wasserstein-distance-based discriminator (W-D). (c) The architecture of the energy-based discriminator (EB-D), giving an ℓ_2 reconstruction error as output. In both of the architectures, Leaky ReLU and batch normalization is employed after every layer except the last.



**Supplementary Material:
Probabilistic Point Cloud
Reconstructions for Vertebral
Shape Analysis**

Supplement for: Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis

Anjany Sekuboyina^{1,2}, Markus Rempfler³, Alexander
Valentinitsch², Maximilian Loeffler², Jan S. Kirschke^{2*}, Bjoern H. Menze^{1*}

¹ Department of Informatics, Technical University of Munich, Germany

² Department of Neuroradiology, Klinikum rechts der Isar, Germany

³ Friedrich Miescher Institute for Biomedical Research, Switzerland

As supplementary content, we present: (1) A detailed description of the complete point-cloud auto-encoder, including the encoder architecture adapted from *point net* [1] (cf. Fig. 6), (2) additional illustrations of point-wise data uncertainty modelled by the proposed σ -VAE (cf. Fig. 7), and (3) Further qualitative results comparing probabilistic reconstructions of healthy and anomalous or fractured vertebrae, along with point-wise Chamfer distance and log-probability between the input and its reconstruction (cf. Fig. 8).

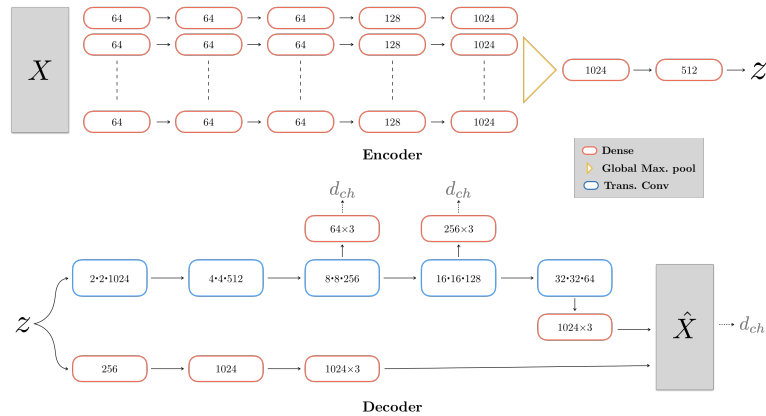


Fig. 6: Architecture for pAE: Architectural details of encoding and decoding paths of the pAE. Note that every layer (except the last, in encoder and in the decoder) is followed by batch normalisation and leaky ReLU. The values in the dense layers indicate the number of nodes while the values in the transposed convolution layers ($- \cdot - \cdot -$ channels) indicate the size of the resulting feature map. For example, the first transposed convolution layer, z is reshaped to $1 \cdot 1 \cdot 64$ and up-convolved to $2 \cdot 2 \cdot 1024$. As the networks operate in point-clouds, they are light-weight. The encoder consists < 3000 parameters. The decoder’s ‘fully-connected’ component has < 4500 parameters. The ‘convolutional-part’ is relatively heavier with $\sim 12\text{M}$ parameters. However, the forward pass takes < 1 second.

* Joint supervising authors.

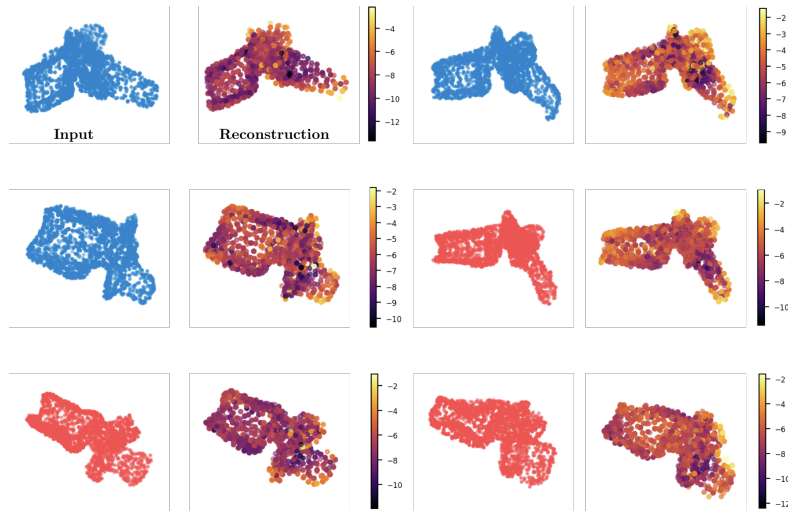


Fig. 7: **Variance modelling** by the proposed σ -VAE for healthy (blue) and fractured (red) cases. Observe a higher variance in the vertebral processes representing the naturally occurring shape variance in a population. Note the lack of high uncertainty values for fractured vertebrae, in accordance with aleatoric uncertainty’s property of capturing only data variance [2]. Hence, predictive variance cannot be used as a means to detect fractures. However, it can reliably be employed as an attenuation factor for improving reconstruction or for computing the reconstruction probability (cf. Fig. 8), thereby enabling fracture detection.

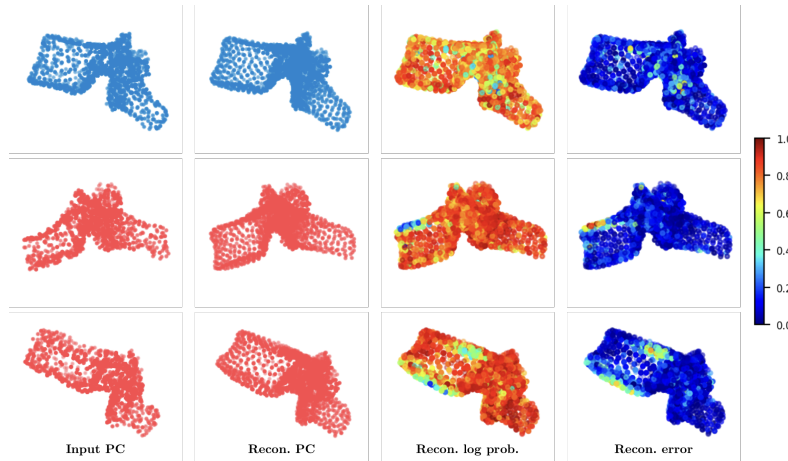


Fig. 8: **Probabilistic reconstructions** of healthy and fractured vertebrae. Alongside spatial localisation of fractures, compare the dynamic range of the reconstruction log probability between healthy and fractured cases when normalised to $[0,1]$. Reconstruction probabilities are relatively uniformly-spread in the healthy case and are pushed to the extremes for a fractured one. This indicates a higher dynamic range in the unnormalised values while reconstructing fractured vertebrae. Thus, a low reconstruction probability (or a high reconstruction error) does indicate an outlier or a fracture.

References

1. Qi, C.R., et al.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. (2017)
2. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS. (2017)



VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images



VERSE: A Vertebrae labelling and segmentation benchmark for multi-detector CT images

Anjany Sekuboyina^{a,b,c,*}, Malek E. Husseini^{a,c}, Amirhossein Bayat^{a,c}, Maximilian Löffler^c, Hans Liebl^c, Hongwei Li^a, Giles Tetteh^a, Jan Kukačka^f, Christian Payer^h, Darko Šternⁱ, Martin Urschler^j, Maodong Chen^k, Dalong Cheng^k, Nikolas Lessmann^l, Yujin Hu^m, Tianfu Wangⁿ, Dong Yang^o, Daguang Xu^o, Felix Ambellan^p, Tamaz Amiranashvili^p, Moritz Ehlke^q, Hans Lamecker^q, Sebastian Lehnert^q, Marilia Lirio^q, Nicolás Pérez de Olaguer^q, Heiko Ramm^q, Manish Sahu^p, Alexander Tack^p, Stefan Zachow^p, Tao Jiang^r, Xinjun Ma^r, Christoph Angerman^s, Xin Wang^{t,u}, Kevin Brown^v, Alexandre Kirszenberg^w, Élodie Puybareau^w, Di Chen^x, Yiwei Bai^x, Brandon H. Rapazzo^x, Timyoas Yeah^A, Amber Zhang^y, Shangliang Xu^z, Feng Hou^B, Zhiqiang He^C, Chan Zeng^D, Zheng Xiangshang^{E,F}, Xu Liming^E, Tucker J. Netherton^G, Raymond P. Mumme^G, Laurence E. Court^G, Zixun Huang^H, Chenhang He^I, Li-Wen Wang^H, Sai Ho Ling^J, Lê Duy Huynh^w, Nicolas Boutry^w, Roman Jakubicek^K, Jiri Chmelik^K, Supriti Mulay^{L,M}, Mohanasankar Sivaprakasam^{L,M}, Johannes C. Paetzold^a, Suprosanna Shit^a, Ivan Ezhov^a, Benedikt Wiestler^c, Ben Glocker^g, Alexander Valentinitzsch^c, Markus Rempfler^e, Björn H. Menze^{a,d,1}, Jan S. Kirschke^{c,1}

^a Department of Informatics, Technical University of Munich, Germany

^b Munich School of BioEngineering, Technical University of Munich, Germany

^c Department of Neuroradiology, Klinikum Rechts der Isar, Germany

^d Department for Quantitative Biomedicine, University of Zurich, Switzerland

^e Friedrich Miescher Institute for Biomedical Engineering, Switzerland

^f Institute of Biological and Medical Imaging, Helmholtz Zentrum München, Germany

^g Department of Computing, Imperial College London, UK

^h Institute of Computer Graphics and Vision, Graz University of Technology, Austria

ⁱ Gottfried Schatz Research Center: Biophysics, Medical University of Graz, Austria

^j School of Computer Science, The University of Auckland, New Zealand

^k Computer Vision Group, iFLYTEK Research South China, China

^l Department of Radiology and Nuclear Medicine, Radboud University Medical Center Nijmegen, The Netherlands

^m Shenzhen Research Institute of Big Data, China

ⁿ School of Biomedical Engineering, Health Science Center, Shenzhen University, China

^o NVIDIA Corporation, USA

^p Zuse Institute Berlin, Germany

^q 1000shapes GmbH, Berlin, Germany

^r Damo Academy, Alibaba Group, China

^s Department of Mathematics, University of Innsbruck, Austria

^t Department of Electronic Engineering, Fudan University, China

^u Department of Radiology, University of North Carolina at Chapel Hill, USA

^v New York University, USA

^w EPITA Research and Development Laboratory (LRDE), France

^x Deep Reasoning AI Inc, USA

^y Technical University of Munich, Germany

^z East China Normal University, China

^A Chinese Academy of Sciences, China

^B Institute of Computing Technology, Chinese Academy of Sciences, China

^C Lenovo Group, China

^D Ping An Technologies, China

^E College of Computer Science and Technology, Zhejiang University, China

^F Real Doctor AI Research Centre, Zhejiang University, China

^G The University of Texas MD Anderson Cancer Center, USA

* Corresponding author.

E-mail address: anjany.sekuboyina@tum.de (A. Sekuboyina).

¹ BM and JSK are supervising authors.

^H Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, China

^I Department of Computing, The Hong Kong Polytechnic University, China

^J The School of Biomedical Engineering, University of Technology Sydney, Australia

^K Department of Biomedical Engineering, Brno University of Technology, Czech Republic

^L Indian Institute of Technology Madras, India

^M Healthcare Technology Innovation Centre, India

ARTICLE INFO

Article history:

Received 5 October 2020

Revised 25 June 2021

Accepted 6 July 2021

Available online 22 July 2021

Keywords:

Spine

Vertebrae

Segmentation

Labelling

ABSTRACT

Vertebral labelling and segmentation are two fundamental tasks in an automated spine processing pipeline. Reliable and accurate processing of spine images is expected to benefit clinical decision support systems for diagnosis, surgery planning, and population-based analysis of spine and bone health. However, designing automated algorithms for spine processing is challenging predominantly due to considerable variations in anatomy and acquisition protocols and due to a severe shortage of publicly available data. Addressing these limitations, the *Large Scale Vertebrae Segmentation Challenge* (VERSE) was organised in conjunction with the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2019 and 2020, with a call for algorithms tackling the labelling and segmentation of vertebrae. Two datasets containing a total of 374 multi-detector CT scans from 355 patients were prepared and 4505 vertebrae have individually been annotated at voxel level by a human-machine hybrid algorithm (<https://osf.io/nqjyw/>, <https://osf.io/t98fz/>). A total of 25 algorithms were benchmarked on these datasets. In this work, we present the results of this evaluation and further investigate the performance variation at the vertebra level, scan level, and different fields of view. We also evaluate the generalisability of the approaches to an implicit domain shift in data by evaluating the top-performing algorithms of one challenge iteration on data from the other iteration. The principal takeaway from VERSE: the performance of an algorithm in labelling and segmenting a spine scan hinges on its ability to correctly identify vertebrae in cases of rare anatomical variations. The VERSE content and code can be accessed at: <https://github.com/anjany/verse>.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The spine is an important part of the musculoskeletal system, sustaining and supporting the body and its organ structure while playing a major role in our mobility and load transfer. It also shields the spinal cord from injuries and mechanical shocks due to impacts. Efforts towards quantification and understanding of the biomechanics of the human spine include quantitative imaging (Löffler et al., 2020a), finite element modelling (FEM) of the vertebrae (Anitha et al., 2020), alignment analysis (Laouissat et al., 2018) of the spine and complex biomechanical models (Oxland, 2016). Biomechanical alterations can cause severe pain and disability in the short term, and can demonstrate worse consequences in the long term, e.g. osteoporosis leads to an 8-fold higher mortality rate (Cauley et al., 2000). In spite of their criticality, spinal pathologies are often under-diagnosed (Howlett et al., 2020; Müller et al., 2008; Williams et al., 2009). This calls for computer-aided assistance for efficient and early detection of such pathologies, enabling prevention or effective treatment.

Vertebral labelling and *vertebral segmentation* are two fundamental tasks in understanding spine image data. Labelled and segmented spines have diagnostic implications for detecting and grading vertebral fractures, estimating the spinal curve, and recognising spinal deformities such as scoliosis and kyphosis. From a non-diagnostic perspective, these tasks enable efficient biomechanical modelling, FEM analysis, and surgical planning for metal insertions. Vertebral labelling can be performed quickly by a medical expert, on smaller datasets, as it follows clear rules (Wigh, 1980). But, manually segmenting them is unfeasible owing to the time required for annotating large structures (e.g. 25 objects of interest with a size of $\sim 10^4$ voxels each). Moreover, the complex morphology of the vertebra's posterior elements combined with lower scan resolutions prevents a consistent and accurate manual delineation. Automating these tasks also involves multiple challenges: highly varying fields of view (FoV) across datasets (unlike brain images), large scan sizes, highly correlating shapes of adjacent vertebrae,

scan noise, different scanner settings, and multiple anomalies or pathologies being present. For example, the presence of vertebral fractures, metal implants, cement, or transitional vertebrae should be considered during algorithm design. Fig. 1 illustrates this diversity using the scans included in the Large Scale Vertebrae Segmentation Challenge (VERSE).

1.1. Terminology

In this section, we introduce three spine-processing terms frequently used in this work: *localisation*, *labelling*, *segmentation*. As used in the rest of the work: *Localisation* is the task of detecting a 3D coordinate on the vertebra and *labelling* is the task of detecting a 3D coordinate on the vertebra as well as identifying the vertebrae. Specifically, labelling supersedes localisation by assigning a 3D coordinate as well as a class to the vertebra (C1-C6, T1-T13, L1-L5, as well as T13 and L6). Unless mentioned otherwise, spine *segmentation* is a voxel-level, multi-class annotation problem, where in each vertebra level has a defined class label (e.g. C1→1, C2→2, T1→8 etc.). It can now be seen that once a vertebra is segmented, its labelling and localisation is implied.

1.2. Prior work

Spine image analysis has received subsistence attention from the medical imaging community over the years. Although computed tomography (CT) is a preferred modality for studying the 'bone' part of a spine due to high bone-to-soft-tissue contrast, there are several prior works on the tasks of labelling and segmenting the spine using multiple modalities in addition to CT such as magnetic resonance imaging (MRI), and 2D radiographs. There are works tackling segmentation (most of which inherently include vertebral labelling), and those tackling labelling specifically from a landmark-detection perspective.

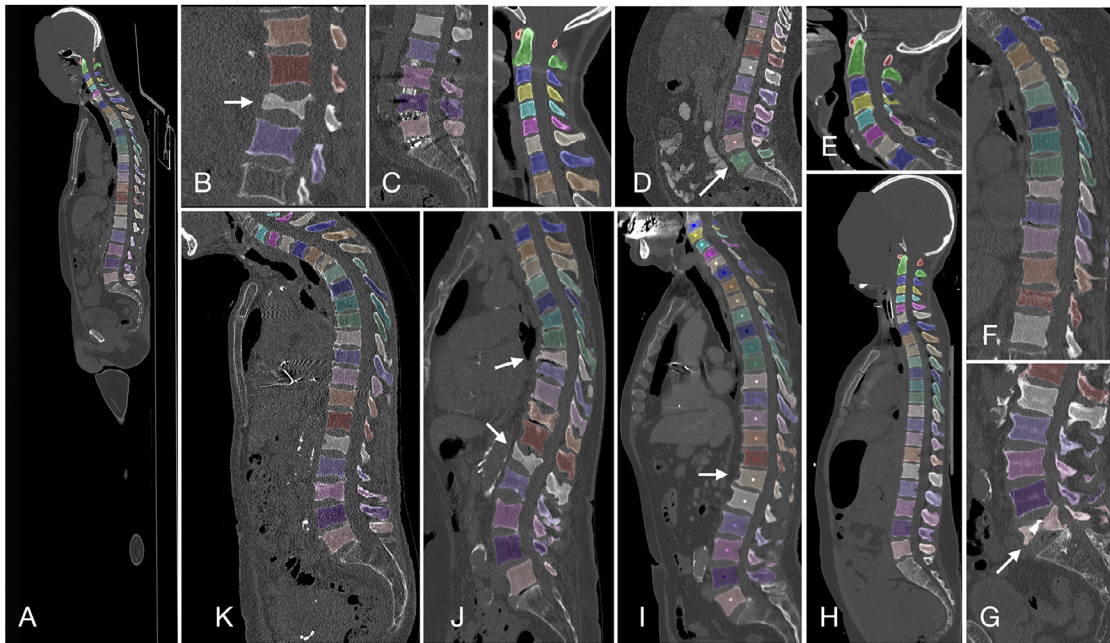


Fig. 1. Example scan slices from the VERSE datasets, labelled clockwise. In addition to the wide variation in the fields of view, we illustrate with fractured vertebrae (B, J), metal insertions (C), cemented vertebrae (G), transitional vertebrae (L6 and T13 in D and I respectively), and a noisy scan (K).

1.2.1. Vertebral segmentation

Traditionally, vertebral segmentation was performed using model-based approaches, which loosely involve fitting a shape prior to the spine and deforming it so that it fits the given spine. The incorporated shape priors range from geometric models (Štern et al., 2011; Ibragimov et al., 2014; 2017), deformed with Markov random fields (MRF) (Kadoury et al., 2011; 2013), statistical shape models (Rasoulian et al., 2013; Pereañez et al., 2015; Castro-Mateos et al., 2015), and active contours (Leventon et al., 2002; Athertya and Kumar, 2016). There are also intensity-based approaches such as level sets (Lim et al., 2014) and *a priori* variational intensity models (Hammernik et al., 2015). Landmark frameworks tackling fully automated vertebral labelling and segmentation from a shape-modelling perspective exist (Klinder et al., 2009; Korez et al., 2015).

With the increased adoption of machine learning in image analysis, works incorporating significant data-based learning components have been proposed. Suzani et al. (2015a) propose using a multi-layer perceptron (MLP) to detect the vertebral bodies and employ deformable registration for segmentation. Similar in philosophy, Chu et al. (2015) propose random forest regression for locating and identifying the vertebrae followed by segmentation performed using random forest classification at a voxel level. Incorporating deep learning, Korez et al. (2016) learn vertebral appearances using 3D convolutional neural networks (CNN) and predict probability maps, which are then used to guide the boundaries of a deformable vertebral model.

The recent advent of deep learning in image analysis and increased computing capabilities have led to works wherein deformable shape modelling and/or vertebral identification was replaced by data-driven learning of the vertebral shape using deep neural networks. Sekuboyina et al. (2017a) perform a patch-based binary segmentation of the spine using a U-Net (Ronneberger et al., 2015) (or a fully convolutional network, FCN) followed by denoising the spine mask using a low-resolution heatmap. Sekuboyina et al. (2017b) propose two neural networks for vertebral segmentation in the lumbar region. First, an MLP learns to regress the localisation of the lumbar region, following which

a U-Net performs multi-class segmentation. Improving on this, Janssens et al. (2018) replace the MLP with a CNN, thus performing multi-class segmentation of lumbar vertebrae with two successive CNNs. Lessmann et al. (2018) propose a two-staged iterative approach, wherein the first stage involves identifying and segmenting one vertebra after another at a lower resolution, followed by a second CNN to refine the lower-resolution masks. Building on this, Lessmann et al. (2019) proposed a single-stage FCN which iteratively regresses the vertebrae's anatomical label and segments it. Once the entire scan is segmented, the vertebral labels are adjusted using a maximum likelihood approach. Approaching the problem from the other end, Payer et al. (2020) propose a coarse-to-fine approach involving three stages, spine localisation, vertebra labelling, and vertebrae segmentation, all three utilising purposefully designed FCNs. Note that (Payer et al., 2020) and (Lessmann et al., 2019) are included in this VERSE benchmark.

1.2.2. Vertebral labelling

Similar to the segmentation works discussed above, classical works on vertebral labelling also involve deformable shape or pose models (Ibragimov et al., 2015; Cai et al., 2015). Learning from data, Major et al. (2013) landmark point using probabilistic boosting trees followed by matching local models using MRFs. As such, works transitioned towards incorporating machine learning using hand-crafted features. Glocker et al. (2012, 2013) employ context features to regress vertebral centroids using regression forests and MRFs. Bromiley et al. (2016) use Haar-like features to identify vertebrae using random forest regression voting. Similarly, Suzani et al. (2015b) employ an MLP to regress the centroid locations. With the incorporation of the ubiquitous CNNs, Chen et al. (2015) proposed a joint-CNN as a combination of random forests for candidate selection followed by a CNN for identifying the vertebrae. Forsberg et al. (2017) employ CNNs to detect the vertebrae followed by labelling them using graphical models.

Going fully convolutional and regressing on input-sized heatmap responses instead of directly learning the centroid locations (which is a highly non-linear mapping), Yang et al. (2017a,b) propose DI2IN, an FCN, for heatmap regression of the

Table 1

Comparing VERSE with other publicly available, annotated CT datasets. In 'Annotations', **L** and **S** refer to annotations concerning the labelling (3D centroid coordinates) and segmentation tasks (voxel-level labels), respectively.

Dataset	#train	#test	Annotations
CSI-Seg 2014 (Yao et al., 2012)	10	10	S
CSI-Label 2014 (Glocker et al., 2012)	242	60	L
Dataset-5 (Ibragimov et al., 2014)	10	–	S (Lumbar)
xVertSeg 2016 (Korez et al., 2015)	15	10	S (Lumbar)
VERSE 2019	80	80	L + S
VERSE 2020	103	216	L + S

vertebral centroids at lower resolution, followed by correction using message passing and recurrent neural networks (RNN) respectively. Utilising a single network termed Btrfly-Net, Sekuboyina et al. (2018, 2020) propose labelling sagittal and coronal maximum intensity projections (MIP) of the spine, reinforced by a prior learnt using a generative adversarial network. Using a three-staged approach, Liao et al. (2018) combine a CNN with a bidirectional-RNN to label and then fine-tune network predictions. Handling close to two hundred landmarks, Mader et al. (2019) use multistage, 3D CNNs to regress heatmaps followed by fine-tuning using regression trees regularised by conditional random fields. Payer et al. (2019) propose a two-stream architecture called spatial-configuration net for integrating global context and local detail in one end-to-end trainable network. With a similar motivation of combining long-range and short-range contextual information, Chen et al. (2019) propose combining a 3D localising network with a 2D labelling network.

1.3. Motivation

Recent spine-processing approaches discussed above are predominantly data-driven, thus requiring annotated data to either learn from (e.g. neural network weights) or to tune and adapt parameters (e.g. active shape model parameters). In spite of this, publicly available data with good-quality annotations is scarce. Eventually, the algorithms are either insufficiently validated or validated in private datasets, preventing a fair comparison. SpineWeb², an archive for multi-modal spine data, lists a total of four CT datasets with voxel-level or vertebra level annotations: CSI2014-Seg (Yao et al., 2012; 2016), xVertSeg (Korez et al., 2015), Dataset-5 (Ibragimov et al., 2014), and CSI2014-Label (Glocker et al., 2012). Table 1 provides an overview of these public datasets. Except Dataset-5, all datasets were released as part of segmentation and labelling challenges organised as part of the computational spine imaging (CSI) workshop at MICCAI. CSI2014-Seg and Label were made publicly available in conjunction with MICCAI 2014 and xVertSeg with MICCAI 2016. Credit is due to these incipient steps towards open-sourcing data, which have yielded interest in spine processing. A significant portion of the work detailed in Section 1.2 is benchmarked on these datasets. However, much is to be desired in terms of *data size* and *data variability*. The largest spine CT dataset with voxel-level annotations to date consists of 25 scans, with lumbar annotations only. CSI-Label, even though it is a collection of 302 scans with high data variability, is collected from a single centre (Department of Radiology, University of Washington), possibly inducing a bias.

With the objective of addressing the need for a large spine CT dataset and to provide a common benchmark for current and future spine-processing algorithms, we prepared a dataset of 374 multi-detector, spine CT (MDCT) scans (an order of magnitude (~20 times) increase from the prior datasets) with vertebral-level

(3D centroids) and voxel-level annotations (segmentation masks). This dataset was made publicly available as part of the *Large Scale Vertebrae Segmentation challenge* (VERSE), organised in conjunction with MICCAI 2019 and 2020. In total, 160 scans were released as part of VERSE'19 and 355 scans for VERSE'20, with a call for fully automated and interactive algorithms for the tasks of *vertebral labelling* and *vertebral segmentation*.

As part of the VERSE challenge, we evaluated twenty-five algorithms (eleven for VERSE'19, thirteen for VERSE'20, and one baseline). This work presents an in-depth analysis of this benchmarking process, in addition to the technical aspects of the challenge. In summary, the contribution of this work includes:

- A brief description of the setup for the VERSE'19 and VERSE'20 challenges (Section 2)
- A summary of the three top-performing algorithms from each iteration of VERSE, along with a description of the in-house, interactive spine processing algorithm utilised to generate the initial annotation. (Section 3)
- Performance overview of the participating algorithms and further experimentation provide additional insights into the algorithms. (Section 4)

2. Materials and challenge setup

2.1. Data and annotations

The entire VERSE dataset consists of 374 CT scans made publicly available after anonymising (including defacing) and obtaining an ethics approval from the institutional review board for the intended use. The data was collected from 355 patients with a mean age of $\sim 59(\pm 17)$ years. The data is multi-site and was acquired using multiple CT scanners, including the four major manufacturers (GE, Siemens, Phillips and Toshiba). Care was taken to compose the data to resemble a typical clinical distribution in terms of FoV, scan settings, and findings. For example, it consists of a variety of FoVs (including cervical, thoraco-lumbar and cervico-thoraco-lumbar scans), a mix of sagittal and isotropic reformations, and cases with vertebral fractures, metallic implants, and foreign materials. Fig. 1 illustrates this variability in the VERSE dataset. Refer to Löffler et al. (2020b); Liebl et al. (2021) for further details on the data composition.

The dataset consists of two types of annotations: 1) 3D coordinate locations of the vertebral centroids and 2) voxel-level labels as segmentation masks. Twenty-six vertebrae (C1 to L5, and the transitional T13 and L6) were considered for annotation with labels from 1 to 24, along with labels 25 and 28 for L6 and T13, respectively. Note that partially visible vertebrae at the top or bottom of the scan (or both) were not annotated. Annotations were generated using a human-hybrid approach. The initial centroids and segmentation masks were generated by an automated algorithm (details in Section 3) and were manually and iteratively refined. Initial refinement was performed by five trained medical students followed by further refinement, rejection, or acceptance by three trained radiologists with a combined experience of 30 years (ML, HL, and JSK). All annotations were finally approved by one radiologist with 19 years of experience in spine imaging (JSK).

2.2. Challenge setup

VERSE was organised in two iterations, first at MICCAI 2019 and then at MICCAI 2020 with a call for algorithms tackling vertebral labelling and segmentation. Both the iterations followed an identical setup, wherein the challenge consisted of three phases: one training and two test phases. In the training stage, participants have access to the scans and their annotations, on which

² spineweb.digitalimaginggroup.ca.

Table 2

Data split and additional details concerning the two iterations of VERSE. Scan split indicates the split of the data into training/PUBLIC test/HIDDEN test phases. Cer, Tho, and Lum refer to the number of vertebrae from the cervical, thoracic, and lumbar regions, respectively. Note that of the 300 patients in VERSE'20, 86 patients are from VERSE'19, resulting in the total patients not being an *ad hoc* sum of the two iterations. VERSE'19 data can be identified by its image ID being less than 500. (Overlap is not absolute owing to the difference in the objectives of the two challenge iterations.).

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

they can propose and train their algorithms. In the first test phase, termed PUBLIC in this work, participants had access to the test scans on which they were supposed to submit the predictions. In the second test phase, termed HIDDEN, participants had no access to any test scans but were requested to submit their code in a docker container. The dockers were evaluated on hidden test data, thus disabling re-training on test data or fine-tuning via overfitting. Information about the data and its split across the two VERSE iterations is tabulated in Table 2. **All the 374 scans of VERSE dataset and their annotations are now publicly available, 2019: <https://osf.io/nqjyw/> and 2020: <https://osf.io/t98fz/>. We have also open-sourced the data processing and the evaluation scripts. All VERSE-content is accessible at <https://github.com/anjany/verse>**

2.3. Evaluation metrics

In this work, we employ four metrics for evaluation, two for the task of labelling and two for the task of segmentation. Note that the evaluation protocol employed for ranking the challenge participants builds on the one presented in this work. Please refer to Appendix A for an overview of the former.

Labelling. To evaluate the labelling performance, we compute the *Identification Rate* (*id.rate*) and localisation distance (d_{mean}): Assuming a given scan contains N annotated vertebrae and denoting the true location of the i^{th} vertebra with x_i and its predicted location with \hat{x}_i , the vertebra i is correctly *identified* if \hat{x}_i is the closest landmark predicted to x_i among $\{x_j \forall j \text{ in } 1, 2, \dots, N\}$ and the Euclidean distance between the ground truth and the prediction is less than 20mm, i.e. $\|\hat{x}_i - x_i\|_2 < 20\text{mm}$. For a given scan, *id.rate* is then defined as the ratio of the correctly identified vertebrae to the total vertebrae present in the scan. Similarly, the localisation distance is computed as $d_{\text{mean}} = (\sum_{i=1}^N \|\hat{x}_i - x_i\|_2) / N$, the mean of the euclidean distances between the ground truth vertebral locations and their predictions, per scan. Typically, we report the mean measure over all the scans in the dataset. Note that our evaluation of the labelling tasks slightly deviates from its definition in (Glocker et al., 2012), where *id.rate* and d_{mean} are computed not at scan-level but at dataset level.

Segmentation. To evaluate the segmentation task, we choose the ubiquitous Dice coefficient (Dice) and Hausdorff distance (*HD*). Denoting the ground truth by T and the algorithmic predictions by P , and indexing the vertebrae with i , we compute the mean Dice score across the vertebrae as follows:

$$\text{Dice}(P, T) = \frac{1}{N} \sum_{i=1}^N \frac{2|P_i \cap T_i|}{|P_i| + |T_i|}. \quad (1)$$

As a surface measure, we compute the mean Hausdorff distance over all vertebrae as:

$$\text{HD}(P, T) = \frac{1}{N} \sum_{i=1}^N \max \left\{ \sup_{p \in P_i} \inf_{t \in T_i} d(p, t), \sup_{t \in T_i} \inf_{p \in P_i} d(p, t) \right\}, \quad (2)$$

where P_i and T_i denote the surfaces extracted from the voxel masks of the i^{th} vertebra and $d(p, t) = \|p - t\|_2$, i.e. a Euclidean distance between the points p and t on the two surfaces.

Outliers. In multi-class labelling and segmentation, there will be cases where the prediction of an algorithm will contain fewer vertebrae than the ground truth. In such cases, d_{mean} and *HD* are not defined for the missing vertebrae. For the sake of analysis in this work, we ignore such vertebrae while computing the averages. This way, we still get a picture of the algorithm's performance on the rest of the correctly predicted vertebrae. The missing vertebrae are anyway clearly penalised by the other two metrics, viz. *id.rate* and *Dice*.

3. Methods

In this section, we present Anduin, our spine processing framework that enabled the medical experts to generate voxel-level annotations at scale. Then, we present details of select participating algorithms.

3.1. Anduin: Semi-automated spine processing framework

Anduin is a semi-automated, interactive processing tool developed in-house, which was employed to generate the *initial* annotations for more than 4000 vertebrae. It is a three-staged pipeline consisting of: 1) *Spine detection*, performed by a light-weight, FCN predicting a low-resolution heatmap over the spine location, 2) *Vertebra labelling*, based on the Btrfly Net (Sekuboyina et al., 2018) architecture working on sagittal and coronal MIPs of the localised spine region, and finally, 3) *Vertebral segmentation*, performed by an improved U-Net (Ronneberger et al., 2015; Roy et al., 2018) to segment vertebral patches, extracted at 1mm resolution, around the centroids predicted by the preceding stage. Fig. 2 gives a schematic of the entire framework. Importantly, the detection and labelling stages offer interaction, wherein the user can alter the bounding box predicted during spine detection as well as the vertebral centroids predicted by the labelling stage. Such *human-in-loop* design enabled the collection of accurate annotations with minimal human effort. We made a web-version of *Anduin* publicly available to the research community that can be accessed at anduin.bonescreen.de. Refer to Appendix B for further details on *Anduin* (at the time of this work) such as network architecture, training scheme, and post-processing steps. Furthermore, without human-interaction, *Anduin* is fully automated. We include this version of *Anduin* in the benchmarking process as 'Sekuboyina A.'. We note that since the ground-truth segmentation masks are generated with *Anduin*-predictions as initialisation, there exists a bias. However, the bias is not as strong for the labelling task as the centroid annotations are sparse and have a high intra- and inter-rater variability.

3.2. Participating methods

Over its two iterations, VERSE has received more than five hundred data download requests. Forty teams uploaded their submissions onto the leaderboards. Of these, eleven and thirteen teams were evaluated for VERSE'19 and VERSE'20, respectively. Table 3 provides a brief synopsis of all the participating teams. Below, we present the algorithms proposed by the best and the second-best-performing teams in each iteration of the challenge. Appendix C provides the details of the remaining algorithms.

Payer C. et al.: *Vertebrae localisation and segmentation with SpatialConfiguration-net and U-net [VERSE'19]*

Vertebrae localisation and segmentation are performed in a three-step approach: spine localisation, vertebrae localisation and

Table 3

Brief summary of the participating methods in VERSE benchmark, ordered alphabetically according to referring author.

	Team / Ref. Author	Method Features	
VERSE'19	 zib / Amiranashvili T.	Multi-stage, shape-based approach. Multi-label segmentation with arbitrary labels for vertebrae. Unique label assignment for based on shape templates. Landmark positions are derived as centres of fitted model.	
	 christoph / Angermann C.	Single-staged, slice-wise approach. One 2.5D U-Net (Angermann et al., 2019) and two 2D U-Nets are employed. The first network generates 2D projections containing 3D information. Then, one 2D U-Net segments the projections, one segments the 2D slices. Labels are obtained as centroids of segmentations.	
	 brown / Brown K.	A 3D bounding box around the vertebra is predicted by regressing on a set of canonical landmarks. Each vertebra is segmented using a residual U-Net and labelled by registering to a common atlas.	
	 iflytek / Chen M.	A three-staged approach. Spine localisation and multi-label segmentation are based on a 3D U-Net. Using the predicted segmentation mask, the third stage employs a RCNN-based architecture to label the vertebrae.	
	 yangd05 / Dong Y.	Single-staged approach. A 3D U-Net based on neural-architecture search is employed to segment vertebrae as 26-class problem. Vertebral-body centre are located using iterative morphological erosion.	
	 huyujin / Hu Y.	Single-staged, patch-based approach. Based on the nnU-Net (Isensee et al., 2019). All three networks are used: a 3D-U-Net at high resolution, a 3D U-Net at low resolution, and a 2D U-Net.	
	 alibabadamo / Jiang T.	Single-staged approach, employing a V-Net (Milletari et al., 2016) backbone with two heads, one for binary- segmentation and the other for vertebral-labelling. Vertebrae C2, C7, T12, and L5 are identified and the rest are inferred from these.	
	 Irde / Kirszenberg A.	Multi-stage, shape-based approach. A combination of three 2D U-Nets generate 3D binary mask of spine. Anchor points on a skeleton obtained from this mask are used for template matching. Five vertebrae are chosen for matching, and one with highest score is chosen as a match.	
	 diag / Lessmann N.	Single-staged, patch-based approach. A 3D U-Net (Lessmann et al., 2019) iteratively identifies and segments the bottom-most visible vertebra in extracted patches, eventually crawling the spine. An additional network is trained to detect first cervical and thoracic vertebrae.	
	 christian_payer / Payer C.	Multi-staged, patch-wise approach. A 3D U-Net regresses a heatmap of the spinal centre line. Individual vertebrae are localized and are identified with the SpatialConfig-Net (Payer et al., 2020). Each vertebra is then independently segmented as a binary segmentation.	
	 init / Wang X.	Multi-staged-approach. A single-shot 2D detector is utilised to localise the spine. A modified Btrfly-Net (Sekuboyina et al., 2018) and a 3D U-Net are employed to address labelling and segmentation respectively.	
	VERSE'20	 deepreasoningai_team1 / Chen D.	Multi-staged, patch-based approach. A 3D U-Net coarsely localises the spine. Then, a U-Net performs binary segmentation, patchwise. Lastly, a 3D Resnet-model identifies the vertebral class taking the vertebral mask and CT-image segmented vertebra.
		 carpediem / Hou F.	Multi-staged approach. First, the spine position is located with 3D U-Net. Second the vertebrae are labelled in the cropped patches. Lastly, U-Net segments individual vertebrae from background using centroids labels.
 poly / Huang Z.		Single-staged, patch-based approach. A U-Net with feature-aggregation and squeeze & excitation module is proposed. Contains two task-specific heads, one for vertebrae labelling and the other for segmentation.	
 Irde / Huỳnh L. D.		A single model with two-stages, a Mask-RCNN-inspired model incorporating RetinaNet is proposed. First stage detects and classifies vertebral Rols. Second stage outputs a binary segmentation for each of the Rols.	
 ubmi / Jakubicek R.		Multi-staged, semi-automated approach (Jakubicek et al., 2020). Stages include: spine-canal tracking, localising and labelling the inter-vertebral disks, and then labelling the vertebrae. Segmentation is based on graph-cuts.	
 htic / Mulay S.		Single-staged approach. A 2D Mask R-CNN with complete IoU loss performs slice-wise segmentation.	
 superpod / Netherton T. J.		Multi-staged approach. Combines a 2D FCN for coarse spinal canal segmentation, a multi-view X-Net (Netherton et al., 2020) for labelling, and a U-Net+ architecture for vertebral segmentation.	
 rigg / Paetzold J.		A naive 2D U-Net performs multi-class segmentation of sagittal slices.	
 christian_payer / Payer C.		Similar to Payer C.'s 2019 submission. Different from it, Markov Random fields are employed for post-processing the localisation stage's output. Additionally, appropriate floating-point optimisation of network weights scans into patches.	
 fakereal / Xiangshang Z.		Both tasks are handled individually. A modified Btrfly-Net (Sekuboyina et al., 2018) detects vertebral key points. An nnU-Net (Isensee et al., 2019) performs multi-class segmentation.	
 sitp / Yeah T.		Two-staged approach containing two 3D U-Nets. First one performs coarse localisation of the spine at low-resolution. Second one performs multi-class segmentation of the vertebra at a higher resolution.	
 aply / Zeng C.		Multi-staged approach. First stage detects five key-points on the spine using a HRNet. Second, improved Spatialconfig-Net (Payer et al., 2019) performs the labelling. Segmentation is now a binary problem.	
 jdlu / Zhang A.		A four-step approach. A patch-based V-Net is used to regress the spine center-line. A key-point localization V-Net predicts potential vertebral candidates. A three-class vertebrae segmentation network obtains main class of each vertebrae. Final labels are obtained using a rule-based postprocessing.	

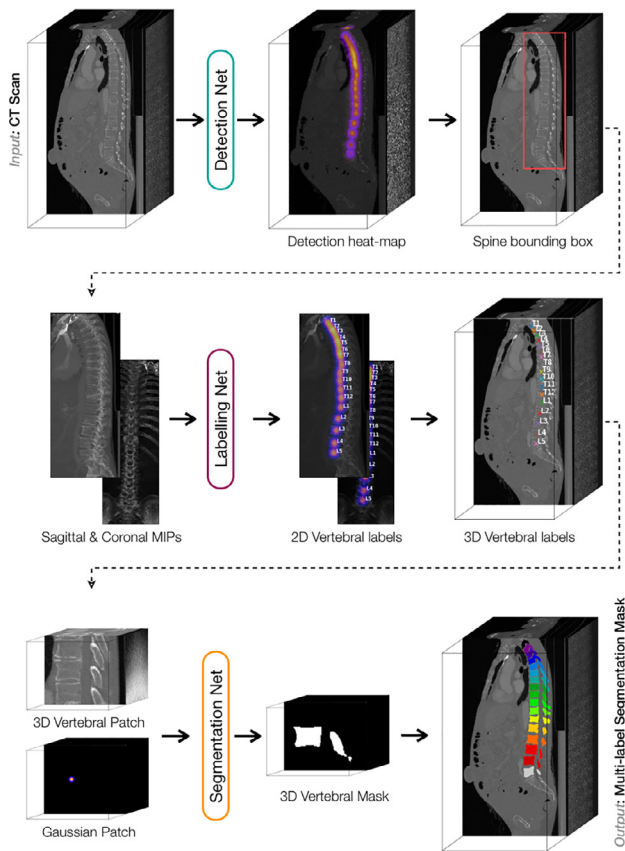


Fig. 2. Our interactive spine-processing pipeline: Schematic of the semi-automated and interactive spine processing pipeline developed in-house. The bold lines indicate automated steps. The dotted lines indicate a possibly interactive step.

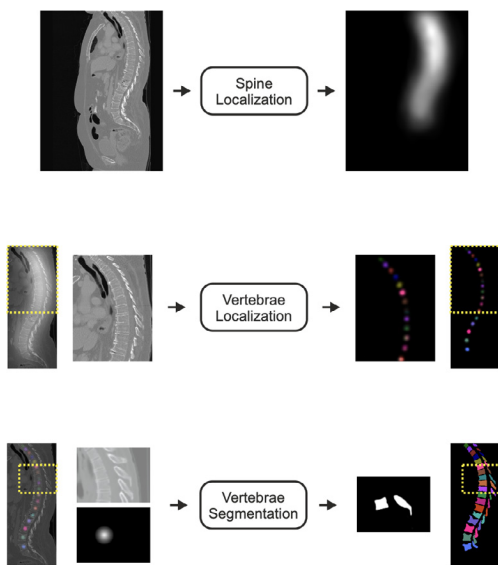


Fig. 3. The three processing stages in Payer C for localisation, identification, and segmentation of vertebrae.

identification, and finally binary segmentation of each located vertebra (cf. Fig. 3). The results of the individually segmented vertebrae are merged into the final multi-label segmentation.

Spine Localisation. To localise the approximate position of the spine, a variant of the U-Net was used to regress a heatmap of the spinal centreline, i.e. the line passing through vertebral centroids, with an ℓ_2 loss. The heatmap of the spinal centreline is generated

by combining Gaussian heatmaps of all individual landmarks. The input image is resampled to a uniform voxel spacing of 8mm and centred at the network input.

Vertebra Localisation & Identification. The SpatialConfiguration-Net (Payer et al., 2020) is employed to localise centres of the vertebral bodies. It effectively combines the local appearance of landmarks with their spatial configuration. Please refer to (Payer et al., 2020) for details on architecture and loss functions. Every input volume is resampled to have a uniform voxel spacing of 2mm, while the network is set up for inputs of size $96 \times 96 \times 128$. As some volumes have a larger extent in the cranio-caudal axis and do not fit into the network, these volumes are processed as follows: During training, sub-volumes are cropped at a random position at the cranio-caudal axis. During inference, volumes are split at the cranio-caudal axis into multiple sub-volumes that overlap for 96 pixels and processed them one after another. Then, the network predictions of the overlapping sub-volumes are merged by taking the maximum response over all predictions.

Final landmark positions are obtained as follows: For each predicted heatmap volume, multiple local heatmap maxima are detected that are above a certain threshold. Then, the first and last vertebrae that are visible on the volume are determined by taking the heatmap with the largest value that is closest to the volume top or bottom, respectively. The final predicted landmark sequence is then the sequence that does not violate the following conditions: consecutive vertebrae may not be closer than 12.5 mm and further away than 50 mm, and a subsequent landmark may not be above a previous one.

Vertebra Segmentation. To create the final vertebrae segmentation, a U-Net is set up with a sigmoid cross-entropy loss for binary segmentation to separate individual vertebrae. The entire spine image is cropped to a region around the localised centroid such that the vertebra is in the centre of the image. Similarly, the heatmap image of the vertebral centroid is also cropped from the prediction of the vertebral localisation network. Both the cropped vertebral image and vertebral heatmap are used as an input for the segmentation network. Both input volumes are resampled to have a uniform voxel spacing of 1 mm. To create the final multi-label segmentation result, the individual predictions of the cropped inputs are resampled back to the original input resolution and translated back to the original position.

Lessmann et al.: Iterative fully convolutional neural networks [VERSE'19]

The proposed approach largely depends on iteratively applied fully convolutional neural networks (Lessmann et al., 2019). Briefly, this method relies on a U-net-like 3D network that analyses a $128 \times 128 \times 128$ region of interest (RoI). In this region, the network segments and labels only the bottom-most visible vertebra and ignores other vertebrae that may be (partly) visible within the RoI. The RoI is iteratively moved over the image by placing it at the centre of the detected piece of the vertebra after each segmentation step. If only part of a vertebra was detected, moving the RoI to the centre of the detected fragment ensures that a larger part of the vertebra becomes visible for the next iteration. Once the entire vertebra is visible in the RoI, the segmentation and labeling results are stored in a memory component. This memory is a binary mask that is an additional input to the network and is used by the network to recognise and ignore already segmented vertebrae. By repeating the process of searching for a piece of vertebra and following this piece until the whole vertebra is visible in the region of interest, all vertebrae are segmented and labeled one after the other. When the end of the scan is reached, the predicted labels of all detected vertebrae are combined in a global maximum likelihood model to determine a plausible labeling for the entire scan, thus

avoiding duplicate labels or gaps. Please refer to (Lessmann et al., 2019) for further details. Note that two publicly available datasets were also used for training: CSI-Seg 2014 (Yao et al., 2012) and the xVertSeg 2016 datasets (Korez et al., 2015). The approach is supplemented with minor changes over (Lessmann et al., 2019) so that: anatomical labelling of the detected vertebra is optimised by minimising a combination of ℓ_1 and ℓ_2 norms; the loss for the segmentation network is a combination of the proposed segmentation error and a cross-entropy loss.

Rib Detection. In order to improve the labeling accuracy, a second network is trained to predict whether a vertebra is a thoracic vertebra or not. As input, this network receives the final image patch in which a vertebra is segmented and the corresponding segmentation mask as a second channel. The network has a simple architecture based on $3 \times 3 \times 3$ convolutions, batch normalisation and max-pooling. The final layer is a dense layer with a sigmoid activation function. At inference time, the first thoracic vertebra and the first cervical vertebra identified by this auxiliary network had a stronger influence on the label voting. Their vote counted three times as much as that of other vertebrae.

Cropping at Inference. Note that if the first visible vertebra is not properly detected, the whole iterative process might fail. Therefore, at inference time, an additional step is added which crops the image along the z-axis in steps of 2.5% from the bottom if no vertebra was found in the entire scan. This helps in case the very first, i.e. bottom-most, vertebra is only visible with a very small fragment. This small element might be too small to be detected as a vertebra but might prevent the network from detecting any vertebra above as the bottom-most vertebra.

Centroid Estimation. Instead of the vertebral centroids provided as training data, the centroids of the segmentation masks were utilised to estimate the “actual” centroids. This was done by estimating the offset between the centroids measured from the segmentation mask (\hat{v}_i) and the expected centroids (v_i). For every vertebra individually, an offset (δ) was determined by minimising $\sum_i \hat{v}_i - v_i + \delta$.

Chen D. et al.: *Vertebrae Segmentation and Localisation via Deep Reasoning [VERSE'20]*

The authors propose deep reasoning approach as a multi-stage scheme. First, a simple U-Net model with a coarse input resolution identifies the approximate location of the entire spine in the CT volume to identify the area of interest. Secondly, another U-Net with a higher resolution is used, zoomed in on the spinal region, to perform binary segmentation on each individual vertebra (bone vs. background). Lastly, a CNN is employed to perform multi-class classification for each segmented vertebra obtained from the second step. The results of the classification and the segmentation are merged into the final multi-class segmentation, which is then used to compute the corresponding centroids for each vertebra.

Spine Localisation. Considering the large volume of whole-body CT scan, the original CT image is down-sampled to a coarse resolution and fed to a shallow 3D-UNet to identify the rough location of the visible spine. The network has the following number of feature maps for both the sequential down and up sampling layers: 8, 16, 32, 64, 128, 64, 32, 16, 8. This is similar to Payer C. et al.'s method for VERSE'19 in Section 3.2. The authors replaced batch normalisation with instance normalisation and ReLU activation with leaky ReLU (leak rate of 0.01), similar to Payer et al. (2020).

Vertebrae Segmentation. The authors train a 3D U-Net model to solely perform binary segmentation (vertebrae bone vs. background) at a resolution of 1mm. Given the natural sequential structure of the vertebrae, inspired by Lessmann et al. (2018), the authors train a model to perform an iterative vertebrae segmentation process along the spine. That is, the model is given the mask of

the previous vertebra and the CT scan as input, and mask for the next vertebrae is predicted. The input is restricted to a small-sized patch obtained from the spine localisation step. A 3D U-Net with the following number of kernels for both the sequential down and up sampling layers is used: 64, 128, 256, 512, 512, 512, 256, 128, 64.

Vertebrae Classification. A 3D ResNet-50 model is used to predict the class of each vertebra. As input, this model takes the segmentation mask obtained in the vertebral segmentation step, as well as the corresponding CT volume, and outputs a single class for the entire vertebrae. Given the prior knowledge of the anatomical structure of the spine and its variations, it can be ensured that the predictions are anatomically valid.

Deep Reasoning Module Given the biological setting of this computer vision challenge, the task is very structured and the proposed models use reasoning to leverage the anatomical structure and prior knowledge. Using the Deep Reasoning framework (Chen et al., 2020), the authors were able to encode and constrain the model to produce results that are anatomically correct in terms of the sequence of vertebrae, as well as only produce vertebral masks that are anatomically possible.

Payer C. et al.: *Improving Coarse to Fine Vertebrae Localisation and Segmentation with SpatialConfiguration-Net and U-Net [VERSE'20]*

The overall setup of the algorithm stays the same as Payer et al.'s approach for VERSE'19 (Payer et al., 2020): a three-stage approach consisting of: spine localisation, vertebrae localisation and identification, and finally binary segmentation of each located vertebra.

This approach, however, differs in its post-processing after the localisation and identification stage, due to an increased variation in the VERSE'20 data. For all vertebrae $i \in \{C1...L6\}$, the authors generate multiple location candidates and identify the ones that maximises the following function of the graph with vertices \mathcal{V} and edges \mathcal{E} modelling an MRF,

$$\sum_{i \in \mathcal{V}} \mathcal{U}(v_i^k) + \sum_{i,j \in \mathcal{E}} \mathcal{P}(v_i^k, v_j^l), \quad (3)$$

where \mathcal{U} describes the unary weight of candidate k of vertebrae i , and \mathcal{P} describes the pairwise weight of the edge from candidate k of vertebrae i to candidate l of vertebrae j . An edge from i to j exists in the graph if v_i and v_j are possible subsequent neighbors in the dataset.

The unary terms are set to the heatmap responses plus a bias, i.e. $u(v_i^k) = \lambda h_i^k + b$, where h_i^k is the heatmap response of the candidate k of vertebra i , b is the bias, and λ is the weighting factor. The pairwise terms penalise deviations from the average vector from vertebrae i to j and are defined as

$$\mathcal{P}(v_i^k, v_j^l) = (1 - \lambda) \left(1 - \left\| \frac{\bar{d}_{i,j} - d_{i,j}^{k,l}}{\|\bar{d}_{i,j}\|_2} \right\|^2 \right), \quad (4)$$

with $\bar{d}_{i,j}$ being the mean vector from vertebra i to j in the ground truth, $d_{i,j}^{k,l}$ being the vector from v_i^k and v_j^l , and $\|\cdot\|$ denoting the Euclidean norm.

The bias is set to 2.0 and also encourages the detection of vertebrae, for which the unary and pairwise terms would be slightly negative. The weighting factor λ set 0.2 to encourage the MRF to more rely on the direction information. For the location candidates of vertex v_i , the authors take the local maxima responses of the predicted heatmap with a heatmap value larger than 0.05. Additionally, as the authors observed that the networks often confuse subsequent vertebrae of the same type, the authors add to the location candidates of a vertebra also the candidates of the previous and following vertebrae of the same type. For these additional

Table 4

Benchmarking VERSE: Overall performance of the submitted algorithms for the tasks of labelling and segmentation over the two test phases. The table reports mean and median (in brackets) measures over the dataset. The teams are ordered according to their Dice scores on the HIDDEN set. Dice and *id.rate* are reported in % and d_{mean} and *HD* in mm. * indicates that the team's algorithm did not predict the vertebral centroids. * indicates a non-functioning docker container. † Jakubicek R. submitted a semi-automated method for PUBLIC and a fully automated docker for HIDDEN.

Team	Labelling				Segmentation			
	Public		Hidden		Public		Hidden	
	<i>id.rate</i>	d_{mean}	<i>id.rate</i>	d_{mean}	Dice	<i>HD</i>	Dice	<i>HD</i>
Payer C.	95.65 (100.0)	4.27 (3.29)	94.25 (100.0)	4.80 (3.37)	90.90 (95.54)	6.35 (4.62)	89.80 (95.47)	7.08 (4.45)
Lessmann N.	89.86 (100.0)	14.12 (13.86)	90.42 (100.0)	7.04 (5.3)	85.08 (94.25)	8.58 (4.62)	85.76 (93.86)	8.20 (5.38)
Chen M.	96.94 (100.0)	4.43 (3.7)	86.73 (100.0)	7.13 (3.81)	93.01 (95.96)	6.39 (4.88)	82.56 (96.5)	9.98 (5.71)
Amiranashvili T.	71.63 (100.0)	11.09 (4.78)	73.32 (100.0)	13.61 (4.92)	67.02 (90.47)	17.35 (8.42)	68.96 (91.41)	17.81 (8.62)
Dong Y.	62.56 (60.0)	18.52 (17.71)	67.21 (71.40)	15.82 (14.18)	76.74 (84.15)	14.09 (11.10)	67.51 (66.05)	26.46 (28.18)
Angermann C.	55.80 (57.19)	44.92 (15.29)	54.85 (57.18)	19.83 (16.79)	43.14 (43.44)	44.27 (35.75)	46.40 (47.98)	41.64 (36.27)
Kirszenberg A.	0.0 (0.0)	155.42 (126.24)	0.0 (0.0)	1000 (1000.0)	13.71 (0.01)	77.48 (86.83)	35.64 (0.09)	65.51 (60.27)
Jiang T.	89.82 (100.0)	7.39 (4.67)	*	*	82.70 (92.62)	11.22 (8.1)	*	*
Wang X.	84.02 (100.0)	12.40 (8.13)	*	*	71.88 (84.65)	24.59 (18.58)	*	*
Brown K.	*	*	*	*	62.69 (85.03)	35.90 (29.58)	*	*
Hu Y.	*	*	*	*	84.07 (91.41)	12.79 (11.66)	81.82 (90.47)	29.94 (20.33)
Sekuboyina A.	89.97 (100.0)	5.17 (3.96)	87.66 (100.0)	6.56 (3.6)	83.06 (90.93)	12.11 (7.56)	83.18 (92.79)	9.94 (7.22)
				VERSE'19				

Team	Labelling				Segmentation			
	PUBLIC		HIDDEN		PUBLIC		HIDDEN	
	<i>id.rate</i>	d_{mean}	<i>id.rate</i>	d_{mean}	Dice	<i>HD</i>	Dice	<i>HD</i>
Chen D.	95.61 (100.0)	1.98 (0.65)	96.58 (100.0)	1.38 (0.59)	91.72 (95.52)	6.14 (4.22)	91.23 (95.21)	7.15 (4.30)
Payer C.	95.06 (100.0)	2.90 (1.62)	92.82 (100.0)	2.91 (1.54)	91.65 (95.72)	5.80 (4.06)	89.71 (95.65)	6.06 (3.94)
Zhang A.	94.93 (100.0)	2.99 (1.49)	96.22 (100.0)	2.59 (1.27)	88.82 (92.90)	7.62 (5.28)	89.36 (92.77)	7.92 (5.52)
Yeah T.	94.97 (100.0)	2.92 (1.38)	94.65 (100.0)	2.93 (1.29)	88.88 (92.93)	9.57 (5.43)	87.91 (92.76)	8.41 (5.91)
Xiangshang Z.	75.45 (92.86)	22.75 (5.88)	82.08 (93.75)	17.09 (4.79)	83.58 (92.69)	15.19 (9.76)	85.07 (93.29)	12.99 (8.44)
Hou F.	88.95 (100.0)	4.85 (1.97)	90.47 (100.0)	4.40 (1.97)	83.99 (90.90)	8.10 4.52	84.92 (94.21)	8.08 (4.56)
Zeng C.	91.47 (100.0)	4.18 (1.95)	92.82 (100.0)	5.16 (2.17)	83.99 (90.90)	9.58 6.14	84.39 (91.97)	8.73 (5.68)
Huang Z.	57.58 (62.5)	19.45 (15.57)	3.44 (0.0)	204.88 (155.75)	80.75 (88.83)	34.06 (27.36)	81.69 (89.85)	15.75 (11.58)
Netherton T.	84.62 (100.0)	4.64 (1.67)	89.08 (100.0)	3.49 (1.6)	75.16 (86.74)	13.56 (6.8)	78.26 (87.44)	14.06 (7.05)
Huynh L.	81.10 (88.23)	10.61 (5.66)	84.94 (90.91)	10.22 (4.93)	62.48 (66.02)	20.29 (16.23)	65.23 (69.75)	20.35 (16.48)
Jakubicek R.†	63.16 (80.0)	17.01 (13.73)	49.54 (56.25)	16.59 (13.87)	73.17 (85.15)	17.26 (12.80)	52.97 (63.56)	20.30 (19.45)
Mulay S.	9.23 (0.0)	191.02 (179.26)	*	*	58.18 (64.96)	99.75 (95.60)	*	*
Paetzold J.	*	*	*	*	10.60 (4.79)	166.55 (265.16)	25.49 24.55	240.61 191.29
Sekuboyina A.	82.68 (93.75)	6.66 (3.87)	86.06 100.0	5.71(3.51)	78.05 (85.09)	10.99 (6.38)	79.52 (85.49)	11.61 (7.76)
				VERSE'20				

candidates from the neighbors, heatmap response is penalised by multiplying it by a factor of 0.1 such that the candidates from the actual landmark are still preferred. Function 3 is solved by creating the graph and finding the shortest negative path from a virtual start to a virtual end vertex.

Another minor change involves usage of mixed-precision networks. The memory consumption of training the networks is drastically reduced due to 16-bit floating-point intermediate outputs, while the accuracy of the networks stays high due to the network weights still being represented as 32-bit floating-point values.

4. Experiments

In this section, we report the performance measures of the participating algorithms in the *labelling* and *segmentation* tasks. Following this, we present a dissected analysis of the algorithms over a series of experiments that help understand the tasks as well as the algorithms.

4.1. Overall performance of the algorithms

The overall performance of the evaluated algorithms for VERSE'19 and VERSE'20 is reported in Tables 4a and 4 b, respectively. We report the mean and the median values of all four evaluation metrics, viz. identification rate (*id.rate*) and localisation distance (d_{mean}) for the labelling task and Dice and Hausdorff distance (*HD*) for segmentation. Note that the algorithms are arranged according to their performance on the corresponding challenge

Table 5

Mean performance (*id.rate* and Dice, in %) of all the evaluated algorithms in both the VERSE iterations. "Top-5" indicates that the mean was computed on the five top-performing algorithms in that year's leaderboard. "All" considers all submitted algorithms.

	VERSE	PUBLIC		HIDDEN	
		All	Top-5	All	Top-5
<i>id.rate</i>	2019	61.4±44.5	83.3±30.7	61.6±43.6	82.4±31.6
	2020	72.5±39.3	93.9±21.0	68.6±42.1	94.4±17.5
Dice	2019	71.2±33.7	82.5±25.9	71.3±32.6	78.9±28.4
	2020	75.2±28.5	89.3±17.9	71.1±32.2	88.8±16.7

leaderboards. Of the evaluated algorithms in VERSE'19, the highest *id.rate* and Dice in the PUBLIC phase were 96.9% and 93.0%, both by Chen M. On the HIDDEN data, these are 94.3% and 89.8%, by Payer C. Similarly, for VERSE'20, Chen D. achieved the highest mean *id.rate* and Dice on both the test phases: 95.6% and 91.7% in PUBLIC and 96.6% and 91.2% in HIDDEN phase. Fig. 4 illustrates the mean and other statistics pertaining to the algorithms' performance as box plots for the four evaluation metrics. Of importance: At least four methods in VERSE'19 achieve a median *id.rate* of 100%. In VERSE'20, this is achieved by seven teams, a majority of the submissions.

Table 5 provides a bigger picture, reporting the mean performance of all the evaluated algorithms as well as the five top-performing algorithms. In 2019, the performance of all methods (incl. Top 5) is consistent between the PUBLIC and HIDDEN phases,

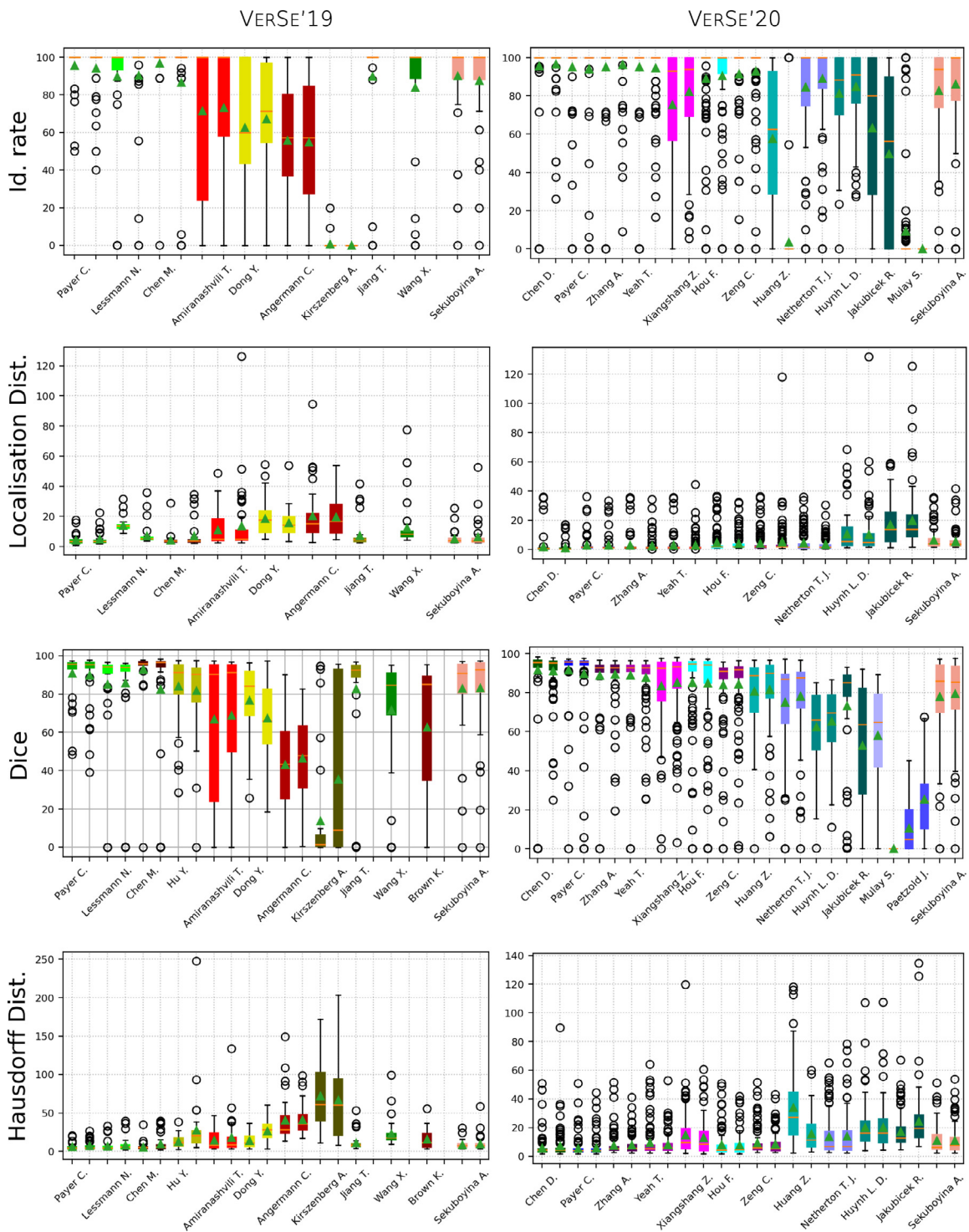


Fig. 4. Overall performance: Box plots comparing all the submissions on the four performance metrics. The plots also show the mean (green triangle) and median (orange line) values of each measure. The two boxes for every team correspond to the performance on the PUBLIC and HIDDEN data. Note that Dice and *id.rate* are on a scale of 0 to 1 while Hausdorff distance (*HD*) and localisation distance (d_{mean}) are plotted in mm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

except for a slight drop in Dice in 2019's HIDDEN phase. However, in 2020, we see that the mean performance of all teams drops, while that of only the top-5 stays relatively consistent. Additionally, observe that the mean *id.rate* and Dice score increased from 2019 to 2020 (for both *All* and *Top-5*). These observations can be attributed to: 1) Supervised algorithms fail to generalise to out-of-distribution cases (L6 in VERSE'19) when their percentage of occurrence in the

dataset is consistent with their low clinical prevalence. 2) With the availability of large, public data with an over-representation of out-of-distribution cases (as in VERSE'20), makes better algorithm design and learning feasible.

In Figs. 5 and 6, we show predictions of the algorithms on the *best*, *median*, and *worst* scans, ranked by the average performance of all the algorithms on every scan. In VERSE'19, the

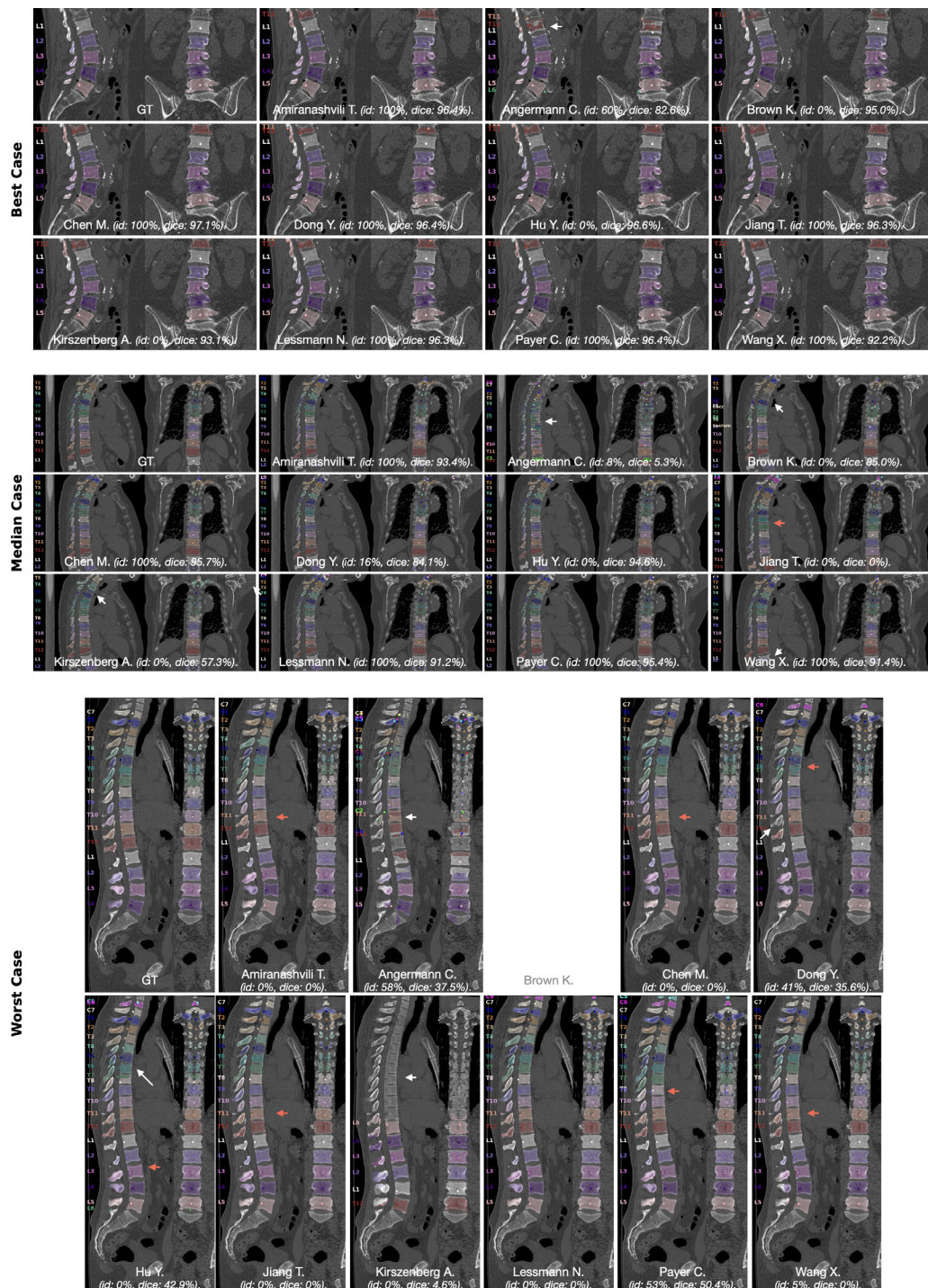


Fig. 5. VERSE'19: Qualitative results of the participating algorithms on the *best*, *median*, and *worst* cases, determined using the mean performance of the algorithms on all cases. We indicate erroneous predictions with arrows. A red arrow indicates mislabelling with a *one-label shift*. From Brown K., the prediction for the worst case was missing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

best scan, a lumbar FoV, is segmented correctly by all the algorithms. The *median* scan, a thoracic FoV with a fracture, is erroneously segmented by a few teams, due to mislabelling (Jiang T., Kirszenberg A., and Wang X.) or stray segmentation (Angermann C., Brown K. and Dong Y.). The *worst*-case scan, interestingly, is an anomalous one, wherein L5 is absent. Seemingly, the lumbar-sacral junction is a strong anatomical pointer for labelling and hence almost every algorithm wrongly labels an L4 as an L5. Medical experts, on the other hand, use the last rib (attached to

T12) to identify the vertebrae and hence would arrive at the correct spine labels. Similarly, in VERSE'20, the *best* case is a lumbar scan. The *median* case is a thoracolumbar scan with severe scoliosis. In spite of this, the majority of the algorithms identify and segment the scan correctly. The *worst* case again occurs due to an anomaly at the lumbar-sacral junction, here due to the presence of a transitional L6 vertebra. Interestingly, the semi-automated approach of Jakubicek R. succeeds in identifying this anomaly correctly.

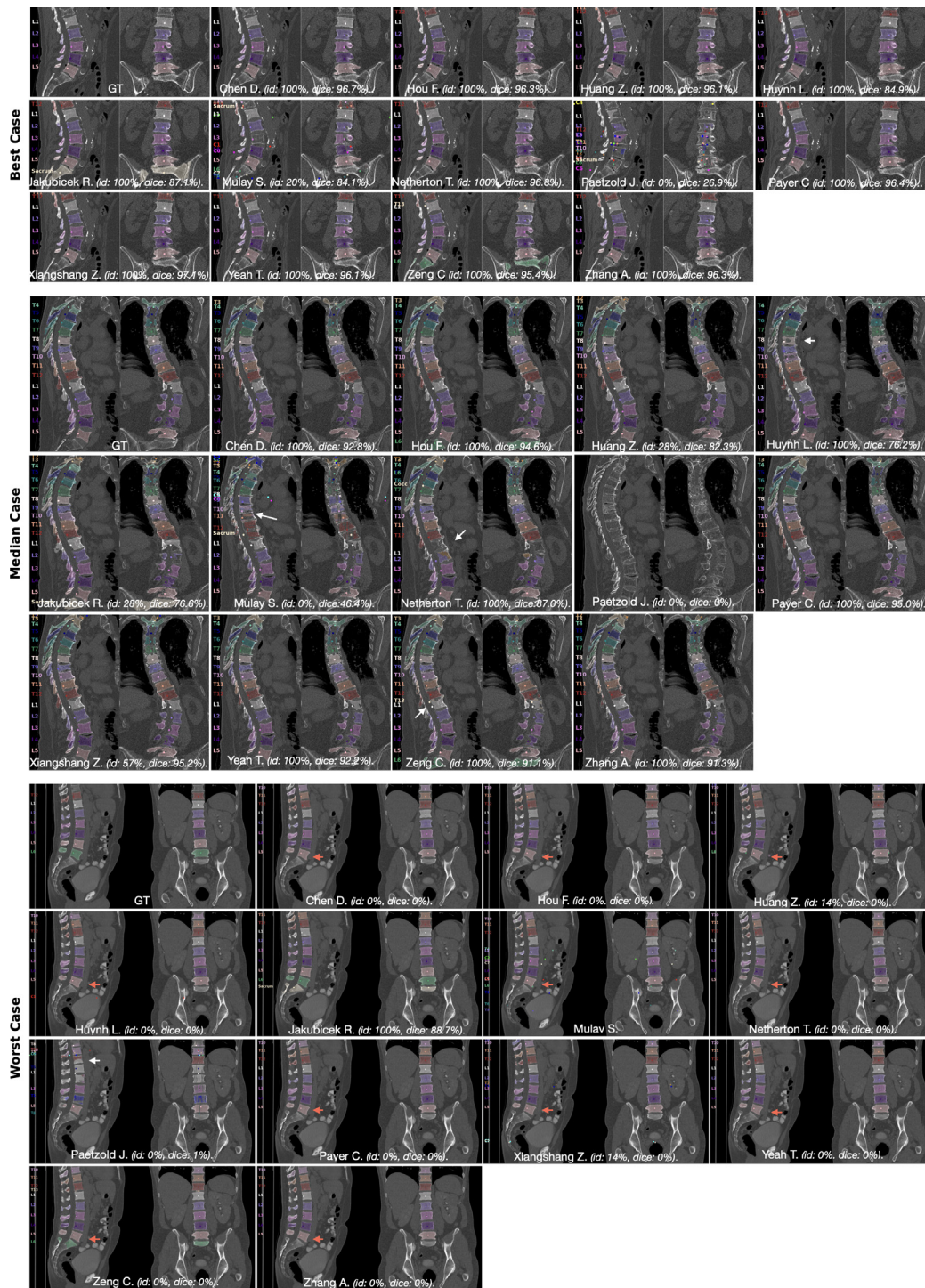


Fig. 6. VERSE'20: Qualitative results of the participating algorithms on the *best*, *median*, and *worst* cases, determined using the mean performance of the algorithms on all cases. We indicate erroneous predictions with arrows. A red arrow indicates mislabelling with a *one-label shift*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Vertebrae-wise and region-wise evaluation

In Fig. 7, we illustrate the mean labelling at segmentation capabilities of the submitted methods at a vertebra-level and region-level (cervical, thoracic, and lumbar).

At a vertebra level, we observe a sudden performance drop in the case of transitional vertebrae (T13 and L6). Concerning L6, None of the methods in VERSE'19 identified the presence of L6. However, in VERSE'20, almost all algorithms identify at least a frac-

tion of the L6 vertebrae. On the other hand, for T13, except for Xiangshang Z., the identification rate widely varies between the PUBLIC and HIDDEN phases for all teams.

Looking at the region-specific performance, VERSE'19 shows a trend of performance-drop in the thoracic region. This could be expected as mid-thoracic vertebrae have a very similar appearance, making them indistinguishable without external anatomical reference. Of course, such a reference (as T12/L1 or C7/T1 junctions) was present in all scans, but apparently not considered by most

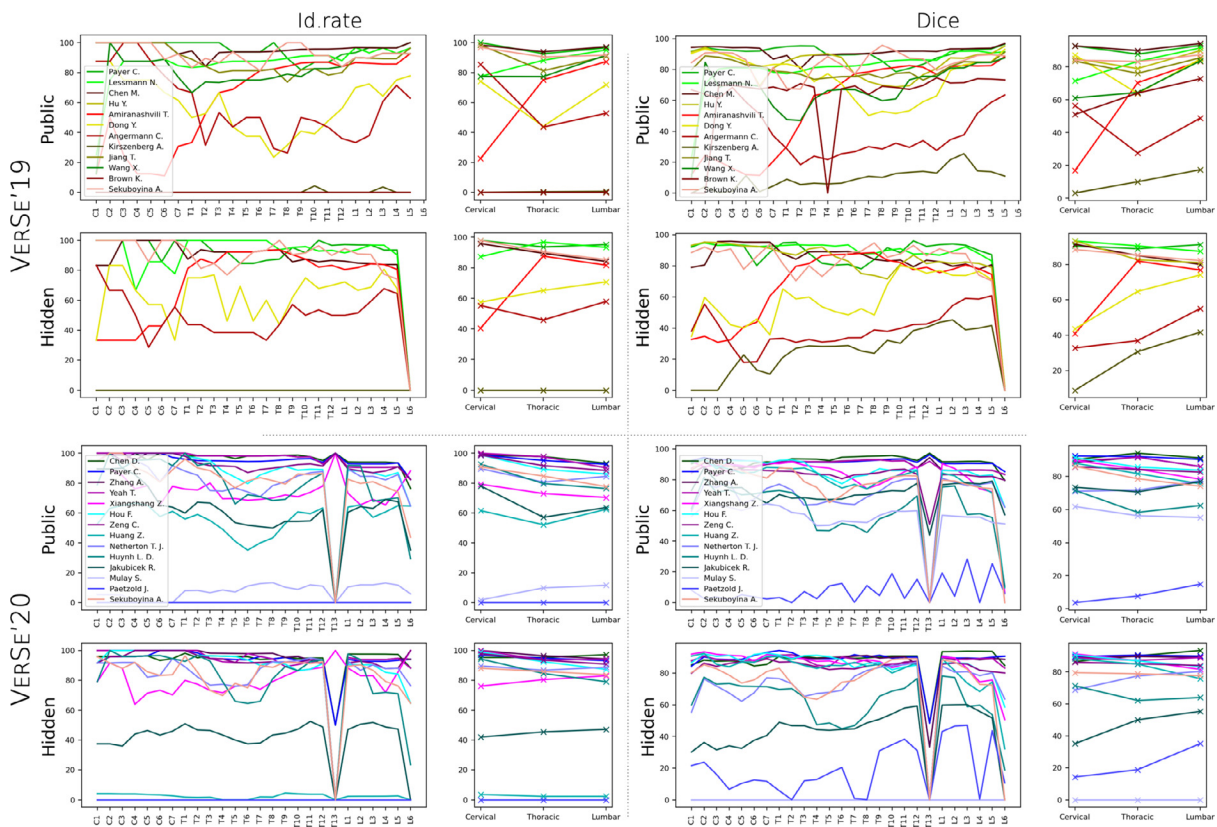


Fig. 7. Vertebra-wise and region-wise performance: Plot shows the mean labelling and segmentation performance of the submitted algorithms at a vertebra level (left) and at a spine-region level (right), viz. cervical, thoracic, and lumbar regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

algorithms. This drop is not observed in VERSE'20. We hypothesize this to be a consequence of better algorithm design because the condition of identifying transitional vertebrae required accurate identification at a local level and reliable aggregation of labels at a global level. We further investigate this behaviour in the following sections.

4.3. Labelling and segmentation at a scan level

When an algorithm is deployed in a clinical setting, minimal manual intervention is desired. Therefore, it is of interest to peruse the *effort* needed for correction. As a proxy, we analyse the number of scans in the dataset that were *successfully* processed. We define *success* using a threshold τ , wherein a scan is said to be *successfully identified* if its *id.rate* is above $\tau_{id.rate}$. Similarly, *successful segmentation* is defined using τ_{Dice} . The fraction of scans successfully processed is denoted by n . In Fig. 8a, we show the behaviour of n at varying thresholds. The best-case scenario for both the tasks is $n = 1, \forall \tau$. The methods in VERSE'20 are closer to this behaviour than VERSE'19, the latter showing more spread over the grid. Especially, Chen D., Payer C., Zhang A., and Yeah T. perfectly identify (*id.rate*=100%) close to 90% of the scans. In 2019, this number was closer to 80% for Chen M., Payer C., and Jiang T. Looking at the Dice curves in 2020, given a vertebra is labelled correctly, its segmentation seems trivial, with the majority of the methods attaining scores of 80–90% on at least 80% of the scans. In 2019, only three methods indicate this performance.

Looking specifically at “failed” scans, we log the number of scans which resulted in less than 5% *id.rate* or Dice in Table 6. When seen in tandem with Fig. 4, this table provides an idea of scan-level failures. Interestingly, in VERSE'20, numerous methods

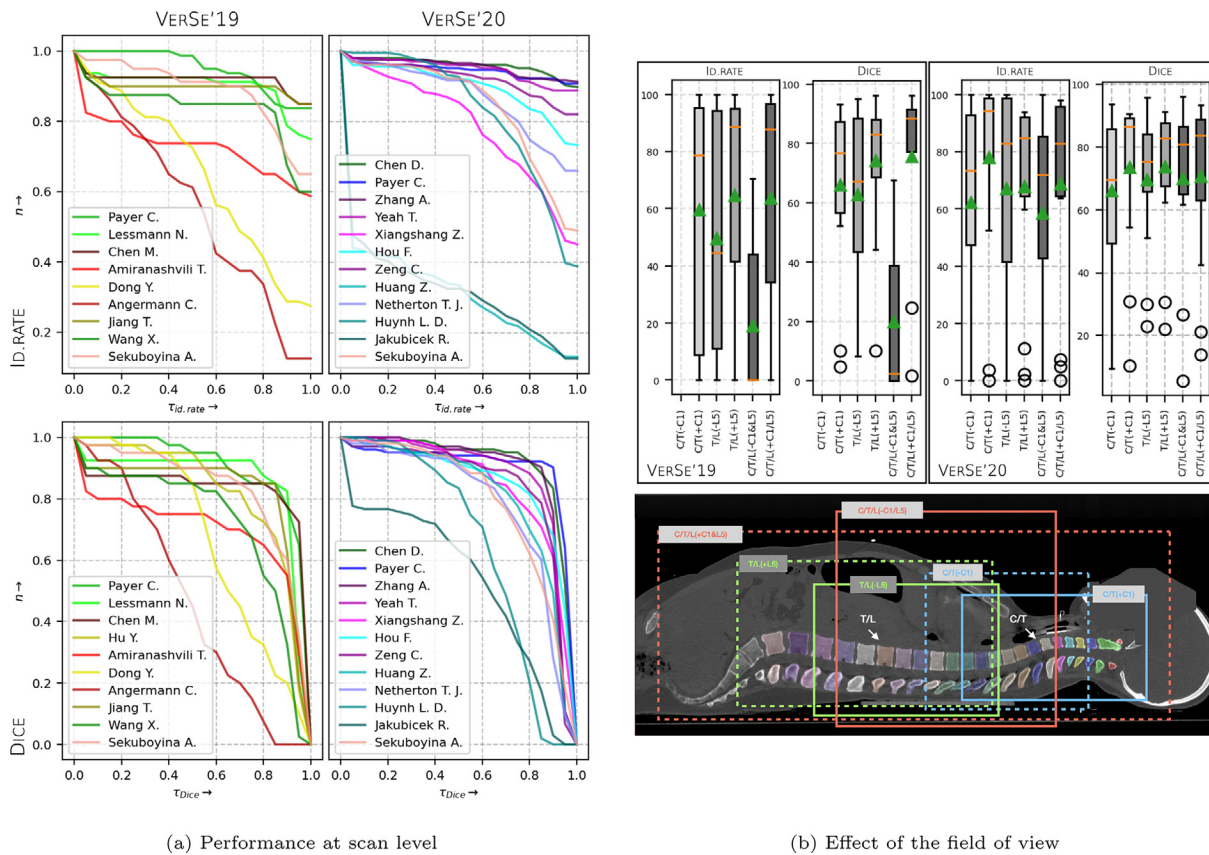
do not show absolute failure in the HIDDEN phase, e.g. Chen D., Zhang A., Yeah T., and Huynh L.

4.4. Effect of field of view on performance

Delving deeper into the region-wise performance of the methods, we ask the question: *What landmark in a scan most aids labelling and segmentation?* For this, we identify four landmarks on the spine: the cranium (if C1 exists), the cervico-thoracic junction (if C7 and T1 coexist), the thoraco-lumbar junction (if T12/T13 and L1 coexist), and lastly the sacrum (if L5 or L6 exists). Based on this, we divide the scans into six categories, namely:

1. $C/T(+C1)$: Cranium and the cervico-thoracic junction are present. Thoraco-lumbar junction absent.
2. $C/T(-C1)$: Cervico-thoracic junction present. Thoraco-lumbar junction absent.
3. $T/L(+L5)$: Sacrum and the thoraco-lumbar junction are present. Cervico-thoracic junction absent.
4. $C/T(-L5)$: Thoraco-lumbar junction present. Sacrum and cervico-thoracic junction absent.
5. $C/T/L(+C1\&L5)$: Full spines. Both cervico-thoracic and thoraco-lumbar junctions are present.
6. $C/T/L(-C1/L5)$: Cervico-thoracic and thoraco-lumbar junctions are present. Either cranium or both cranium and sacrum are absent. (VERSE did not contain any scan with cranium and without sacrum)

Note that in the categories above, L5 refers to the last lumbar vertebra, which could be L4 or L6 as well. Fig. 8b shows an example of a full spine scan with crops that would fall into one of these categories. Once every scan in the dataset is assigned the appropriate category, we compute the mean identification rate and Dice



(a) Performance at scan level

(b) Effect of the field of view

Fig. 8. (a) Fraction of scans, n , with an $id.rate$ or Dice higher than a threshold, τ . The fraction is computed over scans in both the test phases. Uninformative dockers with lines hugging the axes are not visualised (Kirszenberg A., Brown K., Muly S., and Paetzold J.). Hu Y. is not included in the $id.rate$ experiment due to missing centroid predictions. (b) Performance measures of scans grouped according to their field of view. Scans are binned into six categories of FoVs. Please refer to Section 4.4 for details. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

score of every method for every category (cf. Fig. 8b). In VERSE'19, we observe that scans with all lumbar vertebra are easier to process compared to cervical ones (T/L or $C/T/L$ with $L5$). For a similar FoV, we see a large drop when cases do not contain $L5$ or $C1$. This shows the reliance of the VERSE'19 methods on the cranium and sacrum. Interestingly, the reliance on $L5$ is not as drastic in VERSE'20 (refer to categories $-C1\&L5$ and $-L5$). However, the cranium seems to still be a strong reference. Essentially, the median segmentation performance (Dice) coefficient of the methods is $\sim 80\%$ in thoracic and lumbar regions for a variety of FoVs, where at least one of the four landmarks mentioned above is visible. Nonetheless, for cervical ($-thoracic$) scans, there is room for improvement for FoVs without the cranium.

4.5. Performance on anatomically rare scans vs. normal scans

As stated earlier, VERSE'20 was rich in rare anatomical anomalies in the form of transitional vertebrae, viz. $T13$ and $L6$. In Fig. 9, we illustrate the difference in performance of the submitted algorithms between a normal scan and a scan with transitional vertebrae. As expected, we observe a superior performance on normal anatomy when compared to that on rare anatomy. The difference in performance, however, is of interest. In PUBLIC, Yeah T., Zhang A., and Zeng C. have a small drop in performance, with the first two approaches showing a better performance on the rare cases compared to the two top performers, Payer C. and Chen D. In HIDDEN, Payer C. does not show any drop in performance, and outperforms the rest on the rare cases. Arguably, algorithms that either show

a stable performance across anatomies or those that identify (and skip processing) a rare case are preferred in a clinical routine.

4.6. Generalisability of the algorithms

Owing to the HIDDEN test phase in both iterations of VERSE, we have access to the docker containers that can be deployed on any spine scan. The only prerequisite for this being that the scan conforms to the Hounsfield scale (as in VERSE data). Exploring the dockers' ability at clinical translation, we deploy three of the top-performing dockers of VERSE'19 on the HIDDEN set of VERSE'20, and vice versa. Table 7 and Fig. 10 report the cross-iteration performance of these dockers.

Recall that the VERSE'20 data has some overlap with VERSE'19. Therefore, the approaches trained on VERSE'20 perform reasonably well on the VERSE'19 data. There is a drop of $\sim 3\%$, which can be attributed a domain shift between the datasets. Note that Payer C. and Zhang A. succeed in identifying $L6$, while none of the methods in 2019 do, owing to the over-representation of $L6$ in VERSE'20. This underpins our motivation for the second VERSE iteration.

On the other hand, the setting of VERSE'19 methods on VERSE'20 data is more interesting. In addition to a domain shift (due to multi-scanner, multi-centre data in 2020), there are also unseen anatomies. Understandably, we see a drop in performance for Lessmann N. and Chen M. Interestingly, the performance drop is not as large for Payer C. This can be attributed to the way these approaches arrive at the final labels. Lessmann N. depends on identifying the last vertebra. In cases with $L6$, this affects the entire

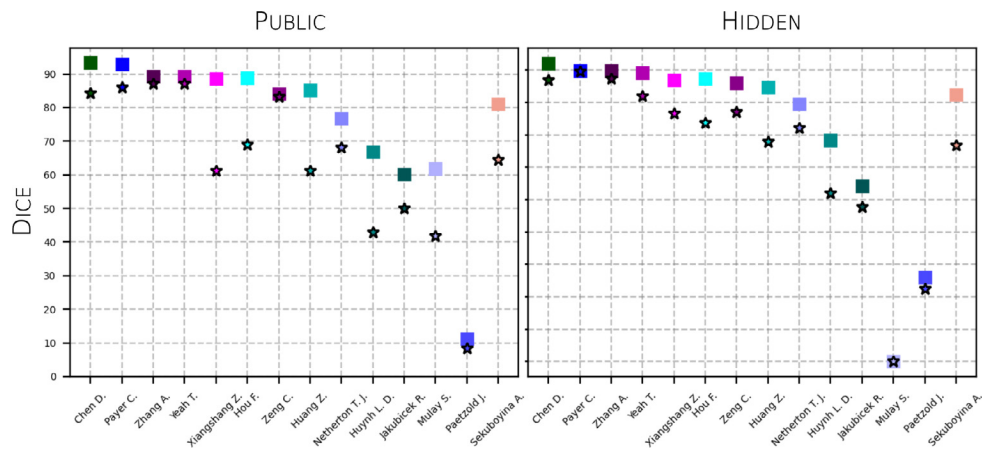


Fig. 9. Performance on transitional vertebrae: Dice scores of the VERSE'20 algorithms computed on anatomically rare scans with transitional vertebrae (*), i.e. T13 and L6, and the normal scans without them (■).

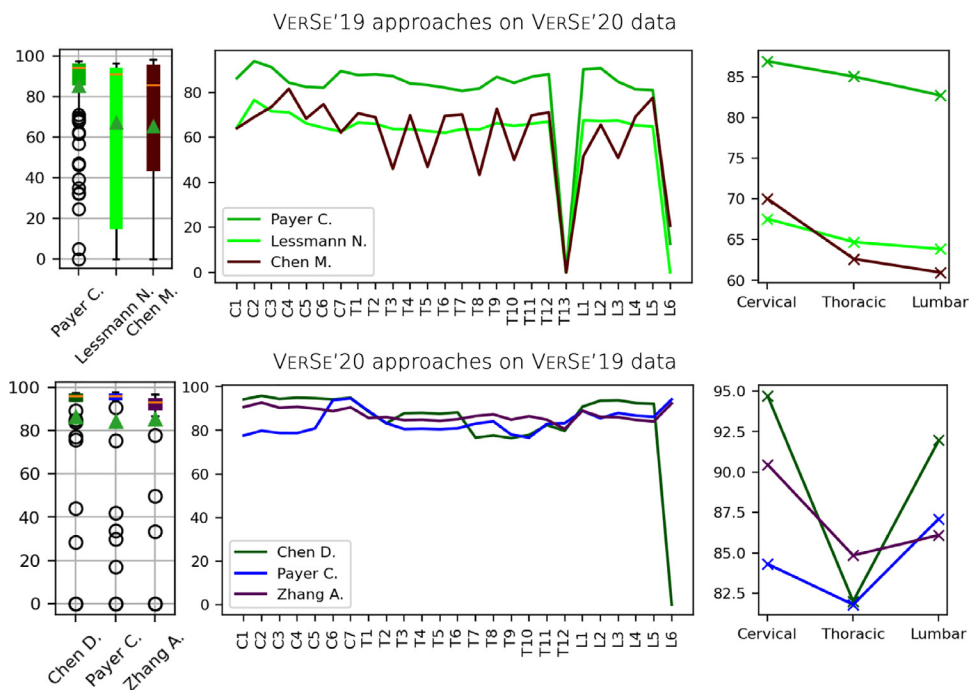


Fig. 10. (Left) Teamwise overall Dice scores of the approaches from one VERSE iteration run on the HIDDEN set of the other iteration. (Center and right) Mean vertebrae-wise, and region-wise Dice scores of the same. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

scan. We assume a similar behaviour for Chen M. In case of Payer C., the presence of L6 was not as detrimental, as the rest of the vertebrae were identified and segmented correctly and the final labels depended prediction confidences during the post-processing stage. Vertebra T13, however, can be ignored due to its absence in VERSE'19.

5. Discussion

5.1. Algorithm design

In this section, we comment on the design of the submitted approaches. Brief descriptions of the evaluated algorithms are provided in Table 3, Section 3, and Appendix C. We look into the following design decisions: pure deep-learning (DL) vs. hybrid models, 3D patch-based vs. 2D slice-wise approach, and a single model vs. a multi-staged approach.

Deep learning vs. hybrid. Out of the twenty-four algorithms benchmarked in this work, twenty-one are purely deep-learning-based, albeit with minor pre- (e.g. intensity-based filtering) and post-processing components (e.g. connected components or morphological operations). Three algorithms: Amiranashvili T., Kirszenberg A, and Jakubicek R. employ statistical shape models. The first two approaches use such models for identifying the vertebrae. The third approach uses it for segmentation using elastic registration. Unlike learning-based approaches, atlases incorporate reliable prior information, thus preventing anatomically implausible results. However, in this benchmark, we see a clear superiority of data-driven, DL approaches compared to the hybrid ones. This is understandable, given the size of VERSE. Better integration of shape-based and learning-based ones is of interest, thus enabling segmentation with anatomical guarantees.

3D patch-based vs. 2D slice-wise segmentation. Common among all the algorithms is the motivation that a clinical spine scan's size is large for current-generation GPU memory. We can

Table 6 Number of scans in each subset of VERSE with an id rate or Dice score less than 5%. Reported values are absolute number of scans from a maximum of: 40 scans each for VERSE'19's PUBLIC and HIDDEN sets, and 103 scans each for VERSE'20's test sets.

		VERSE'20																																											
		VERSE'19				VERSE'20				VERSE'20				VERSE'20																															
		Lessmann		Amitanashvili		Angermann Kirszenberg		Jiang T.		Wang X.		Brown K.		Hu Y.		Sekuboyina A.		Chen D.		Payer C.		Zhang A.		Yeah T.		Xiangshang Z.		Hou F.		Zeng C.		Huang Z.		Netherton T.		Huyh L.		Jafallicek R.		Mulay S.		Paetzold J.		Sekuboyina A.	
		Payer C.	N.	Chen M.	T.	Dong Y.	C.	A.	Jiang T.	Wang X.	Brown K.	Hu Y.	A.	Chen D.	Payer C.	Zhang A.	Yeah T.	Xiangshang Z.	Hou F.	Zeng C.	Huang Z.	Netherton T.	Huyh L.	Jafallicek R.	Mulay S.	Paetzold J.	Sekuboyina A.	Chen D.	Payer C.	Zhang A.	Yeah T.	Xiangshang Z.	Hou F.	Zeng C.	Huang Z.	Netherton T.	Huyh L.	Jafallicek R.	Mulay S.	Paetzold J.	Sekuboyina A.				
PUBLIC	id.rate	0	3	1	6	3	2	38	3	3	-	-	1	4	3	4	4	5	5	4	16	4	1	77	82	-	3	3	2	3	3	3	1	4	1	4	1	77	82	-	3				
	Dice	0	3	1	7	0	2	28	4	4	8	0	1	3	2	3	3	1	4	4	3	1	4	1	16	6	52	2	2	2	2	2	2	2	2	1	4	1	16	6	52	2			
HIDDEN	id.rate	0	2	4	8	1	4	40	-	-	-	-	1	0	3	0	0	0	3	2	99	2	0	31	-	-	3	3	0	0	0	0	0	2	0	0	31	-	-	6	6	-	3		
	Dice	0	3	5	7	0	1	14	-	-	-	-	1	0	3	0	0	1	3	3	0	2	0	23	-	-	6	6	0	0	0	0	2	0	0	23	-	-	6	6	-	1			

Table 7

Mean Dice (%) of running the three of the top-performing dockers of one VERSE iteration on HIDDEN set of the other iteration.

V'19 approaches on V'20 data	
Payer C.	85.21
Lessmann N.	66.96
Chen M.	65.21
V'20 approaches on V'19 data	
Chen D.	86.44
Payer C.	84.11
Zhang A.	85.42

draw two lines of algorithms among the benchmarked ones: First, those performing 2D slice-wise segmentation (e.g. Angermann C., Kirszenberg A., Mulay S., Paetzold J.). Second, which form the majority, are the approaches that perform patch-wise segmentation in 3D using architectures such as 3D U-Net (Çiçek et al., 2016), V-Net (Milletari et al., 2016), or nnU-Net (Isensee et al., 2019). The second category can further be split into approaches performing multi-label segmentation, and those performing binary segmentation.

Observe that, in general, 3D processing is preferable naive 2D slice-wise segmentation. More so, when compared to 2D slice-wise multi-label segmentation. This is expected because slice-wise processing, in spite of offering a larger FoV and memory efficiency, ignores crucial 3D context for an anatomically large structure such as a spine. Moreover, labelling the vertebrae becomes noisy as not every vertebra is visible in every slice.

Single model vs. multi-staged. One principal categorisation of the benchmarked algorithms is into two categories based on the number of stages they employ to tackle the tasks of labelling and segmentation, as demonstrated by some representative algorithms listed below:

1. Single-stage: Lessmann N., Jiang T., Huang Z., and Huynh D.
2. Multi-staged: Chen D., Payer C., Zhang A., and Netherton T.

Typically, single-staged models work with 3D patches. The likes of Lessmann N. perform iterative identification and segmentation and determine a label arrangement using maximum likelihood estimation. Jiang T. and Huang Z. propose dedicated architectures with multiple heads, one each for the labelling and segmentation tasks, thus exploiting their interdependency. nnU-Net or 3D-UNet-based multi-label classification followed by final labelling is also a recurring theme.

On the other hand, numerous sequential frameworks have also been proposed. Payer C., for instance, perform labelling and segmentation in three stages of localisation, then labelling, and finally binary vertebral segmentation. Zhang A. propose a four-stage approach involving spine-centerline detection, vertebral candidate prediction, and a three-class segmentation of the localised spine. Following this, final labels are identified based on certain spine-centric rules.

As evidenced by the performance, one cannot propose a 'winner' among the two categories. Both categories equally span the upper regions of the leaderboards. The first category could possibly result in numerous inferences of large patches per scan (resulting in longer inference times), while the second approach could be prone to errors compounding from a preliminary stage of the sequence.

5.2. On rare anatomical variations: Transitional vertebrae

VERSE'19 included two cases with L6 in the train set, a proportion resembling its clinical occurrence. We observed that almost

every algorithm fails to segment the one L6 in the HIDDEN set. A major motivation for the second iteration of VERSE, was hence, to increase number of anatomically anomalous cases. VERSE'20 included six cases with T13 (2/2/2 in TRAIN/PUBLIC/HIDDEN) and 47 cases with an L6 (15/15/17). The effect of this increase in transitional vertebrae can be seen in Fig. 7, with L6 now being detected and segmented, at least in some cases. Surprisingly, T13, if occurring only twice is successfully identified by some methods. Note that Xiangshang Z. is the only approach that successfully identifies all T13 instances in both test phases.

This contradictory behaviour of better performance of approaches in the case of T13 compared to L6, in spite of higher numbers gives us some insights into the task at hand. For T13, the sequence of vertebral labels gives a strong prior. In the case of L6, which itself acts as a strong prior due to the sacrum, its reliable detection doesn't seem as consistent. Hanaoka et al. (2017), for example, recognise this issue and work towards directly predicting such abnormal numbers. Nonetheless, the improved behaviour of the approaches in such anatomical variations brings us closer to realising automated algorithms in clinical settings.

5.3. Limitations of our study

The scale, clinical similitude, data and anatomical variability are the strengths of the VERSE benchmark. In this section, we identify some limitations of this study.

Foremost among the limitations is the lack of inter-rater annotations. Owing to the effort involved in creating the voxel-level annotations for a multitude of vertebrae, the hierarchical process of okaying an annotation, and the use of a machine in the annotation process, the decision of having multiple-raters was delegated to future challenge iterations. This would eventually enable algorithms to predict uncertainty, inter-rater variability studies, and learning annotator biases.

Putting aside the insufficiency of the Dice metric for evaluating segmentation performance (Taha and Hanbury, 2015), the metrics in the spine literature have a major short-coming: one-label shift, where the labels of the predicted mask are off by one label (cf. Fig. 6, Worst Case). One-label shift penalises the current metrics more than label mixing, which results in unusable masks. The drastic drop in performance of Chen M. between the PUBLIC and HIDDEN phases (Table 4a) was due to this issue. Therefore, research towards better domain-specific evaluation metrics is of interest, more so for differentiable variants enabling neural network optimisation.

6. Conclusions

The Large Scale Vertebrae Segmentation Challenge (VERSE) was organised in two iterations in conjunction with MICCAI 2019 and 2020. VERSE, publicly made available 374 CT scans from 355 patients, the largest spine dataset to date with accurate centroid and voxel-level annotations. On this data, twenty-five algorithms (twenty-four participating algorithms, one baseline) are evaluated for the tasks of vertebral labelling and segmentation. This work describes the challenge setup, summarises the baseline and the participating algorithms, and benchmarks them with each other. The best algorithm in terms of mean performance in VERSE'19 achieves identification rate of 94.25% and a Dice score of 89.80% (Payer C.) on the HIDDEN test set. In VERSE'20, these numbers are 96.6% (*id.rate*) and 91.72% (Dice), achieved by Chen D. Based on the statistical ranking method chosen for evaluating VERSE challenges, Payer C.'s approach led the leaderboard due to its better and relatively consistent performance on healthy as well as the anatomically rare cases.

Aimed at understanding the algorithms' behaviour, we present an in-depth analysis in terms of the spine region, fields of view, and manual effort. We make the following key observations: (1) The performance of algorithms, on average, increased from VERSE'19 to VERSE'20, in spite of the data being more multi-centred and anomalous, (2) Spine processing, for now, is better approached in 3D, either as large patches or in a appropriately designed sequence of stages, and (3) Transitional vertebrae (T13 and L6) can be efficiently handled given sufficient data and post-processing. We hope that the VERSE dataset and benchmark will enable researchers to contribute towards more accurate and reliable clinical translation of their spine algorithms.

As stated, future directions could include the incorporation of multi-raters, inter-rater variability, and spine-centred evaluation measures. Additionally, modelling the sacrum is of interest for load analysis. Lastly, in spite of labelling and segmentation being inter-dependent, our motivation for having two tasks was to enable participation in individual tasks. However, our experience shows this to be redundant. Moreover, the VERSE challenges did not explicitly require the participating algorithms to be optimised for run time. Including this as an objective could bring in added insights into algorithm design. We bring these observations to the attention of future attempts at benchmarking.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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). We acknowledge NVIDIA Corporation's support with the donation of the GPUs used for this research.

FA and TA from Zuse Intitute Berlin are partly funded by the German Ministry of Research and Education (BMBF) Project Grant 3FO18501 (Forschungscampus MODAL).

Appendix A. Challenge Evaluation and Ranking

A1. Statistical tests & points

In this technical report (Section 4), we reported four performance measures, *id.rate*, d_{mean} , Dice, and *HD*. However, recall from 2, that *HD* and d_{mean} are undefined in the case of missing vertebral predictions. Therefore, to rank the teams in VERSE'19, the missing predictions for vertebrae were substituted by a maximum Euclidean distance of 1000mm for d_{mean} and 100mm for *HD*. Expecting more missed predictions in VERSE'20, and in order to avoid inducing a bias due to such substitution, *HD* and d_{mean} were not used to rank the teams in VERSE'20.

Once computing the performance measures, we compare them. Inspired by Maier-Hein et al. (2018) and Menze et al. (2014), the comparison and ranking of the participating algorithms were based on a scheme based on statistical significance. The value of the performance measure obtained from each scan in the cohort was treated as a sample from a distribution and the Wilcoxon signed-rank test with a 'greater' or 'less' hypotheses testing (as appropriate for the performance metric) was employed to test the significance of the difference in performance between a pair of participants. A p -value of 0.001 was chosen as the threshold to ascertain a significant difference. Following this, a *point* was assigned to the better team. All possible pairwise comparisons were performed for

every performance measure. Each comparison awards a point to a certain team unless the difference is not statistically significant. For every measure, the points are aggregated at a team level and normalised with the total number of participating teams in the experiment to obtain a score between 0 and 1.

Lastly, for every team, the normalised points across the measures are combined as described in the next section, which describes particulars of point-computation for the ranking pertaining to the challenge.

The points scored by each team are reported in Tables A.8a and A.8 b respectively. Illustrated in Figs. A.11 and A.12 are the p -values of the significance as well as their binarised versions (thresholded at $p = .001$) that ensue from the pairwise comparisons.

A2. Final ranking: Combining all the scores

Fig. A.13 illustrates how the performance of the algorithms over the multiple stages was combined to construct one ranking scheme. Tables A.8a and A.8 b also report the normalised points. The rationale in choosing this presented scheme was as follows:

- d_{mean} and HD , compared to $id.\text{rate}$ and Dice, are weighted at a ratio of 1 : 2 in order to de-emphasise the contribution of the upper bounds chosen on the former measures in case of missing predictions. (This does not apply to VERSE'20.)
- HIDDEN has twice the weight as PUBLIC as it was evaluated on a completely hidden dataset, thus nullifying the chance of overfitting or retraining on the test set.
- Lastly, the segmentation task has twice the weight of the labelling task as the latter can possibly be a consequence of the former, as was the final goal of this challenge.

Appendix B. Description of Anduin

The *Anduin* framework was used to assist the data team in the creation of the ground truth. Please refer to Löffler et al. (2020b) for an overview of annotation-creation for VERSE. Given the CT scan of a spine, our framework aims to predict accurate voxel-level segmentation of the vertebrae by splitting the task in to three sub-tasks: spine detection, vertebrae labelling, and vertebrae segmentation. In the following section, the network architectures, loss functions, and training and inference details of each of these modules is elaborated. Fig. 2 gives an overview of the proposed framework and Fig. B.14 details the architectures of the networks employed in the three sub-tasks.

B1. Notation.

The input CT scan is denoted by $\mathbf{x} \in \mathbb{R}^{h \times w \times d}$ where h , w , and d are the height, width, and depth of the scan respectively. The annotations available are, (1) the vertebral centroids, denoted by $\{\mu_i \in \mathbb{R}^3\}$ for $i \in \{1, 2, \dots, N\}$. These are used to construct the ground truth for the detection and labelling tasks, denoted by \mathbf{y}_d and \mathbf{y}_l , respectively. (2) the multi-label segmentation masks, denoted by $\mathbf{y}_s \in \mathbb{Z}^{h \times w \times d}$.

B2. Spine detection

To detect the spine, we propose a parametrically-light, 3D, FCN operating at an isotropic resolution of 4mm. This network regresses a 3D volume consisting of Gaussians at the vertebral locations as shown in Fig. B.14. The Gaussian heatmap is generated

Table A.8

Point counts of the submitted approaches of (a) VERSE'19 and (b) VERSE'20, based on the proposed pairwise, statistical comparison. * indicates a non-functioning docker container. † Jakubicek R. submitted a semi-automated method for PUBLIC and a fully-automated docker for HIDDEN.

Team	Normalised Points	Labelling				Segmentation			
		Public		Hidden		Public		Hidden	
		$id.\text{rate}$	d_{mean}	$id.\text{rate}$	d_{mean}	Dice	HD	Dice	HD
VERSE'19 (a)									
Payer C.	0.691	3	7	3	5	8	8	5	5
Chen M.	0.597	5	7	2	4	10	8	3	4
Lessmann N.	0.496	3	1	4	3	4	5	3	5
Hu Y.	0.279	*	*	*	*	4	4	3	3
Dong Y.	0.216	1	1	1	1	2	4	2	1
Amiranashvili T.	0.215	1	1	1	1	1	3	2	2
Jiang T.	0.140	3	5	*	*	4	4	*	*
Angermann C.	0.107	1	1	1	1	1	2	0	1
Wang X.	0.084	2	3	*	*	2	3	*	*
Brown K.	0.022	*	*	*	*	1	1	*	*
Kirszenberg A.	0.007	0	0	0	0	0	1	0	0
Team	Normalised Points	Labelling				Segmentation			
		Public	Hidden	Public	Hidden	Public	Hidden		
		$id.\text{rate}$	$id.\text{rate}$	Dice	Dice	Dice	Dice		
VERSE'20 (b)									
Payer C.	0.675			6	4	11	10		
Chen D.	0.581			7	5	10	7		
Yeah T.	0.453			6	5	7	5		
Zhang A.	0.453			6	5	7	5		
Hou F.	0.393			5	4	7	4		
Zeng C.	0.333			6	4	5	3		
Xiangshang Z.	0.316			2	2	6	4		
Netherton T.	0.222			3	3	3	2		
Huang Z.	0.171			1	0	4	2		
Huynh L.	0.119			3	2	1	2		
Jakubicek R.†	0.085			1	1	3	0		
Mulay S.	0.017			0	*	1	*		
Paetzold J.	0.0			*	*	0	0		

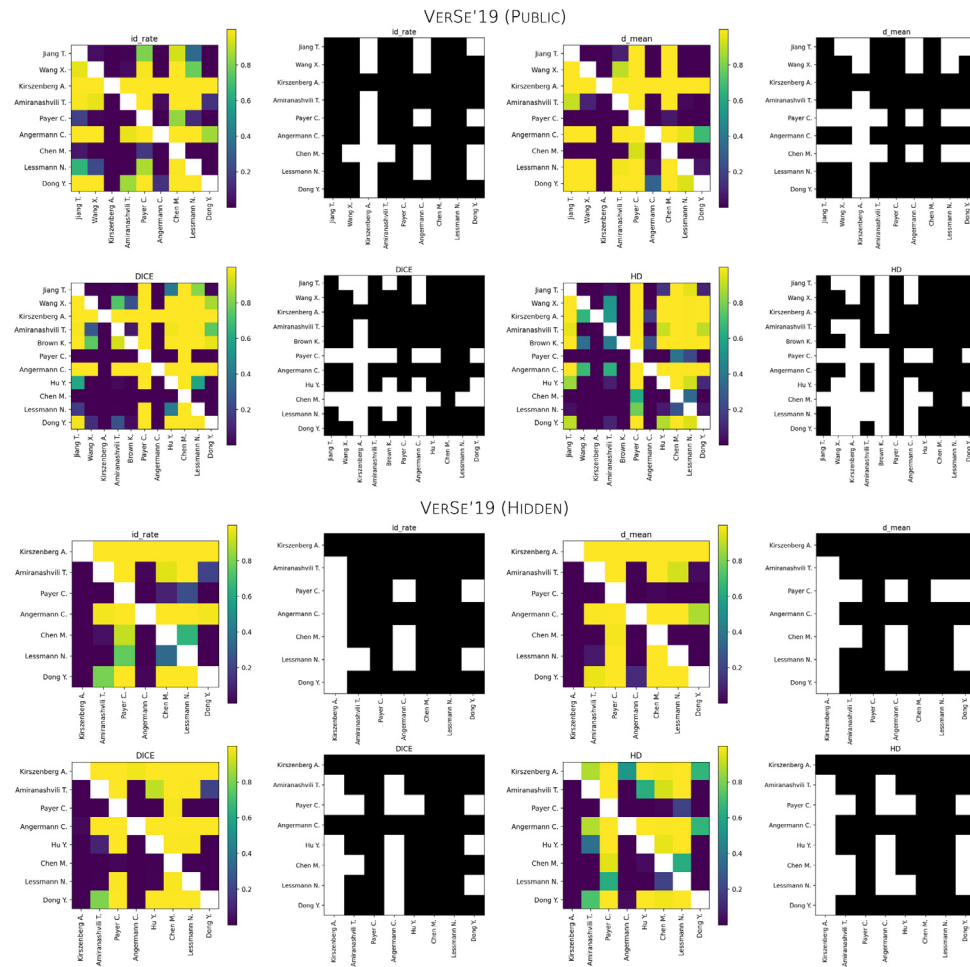


Fig. A.11. VERSE'19 points: Illustrating the p -value matrices and their binarised versions for every metric used.

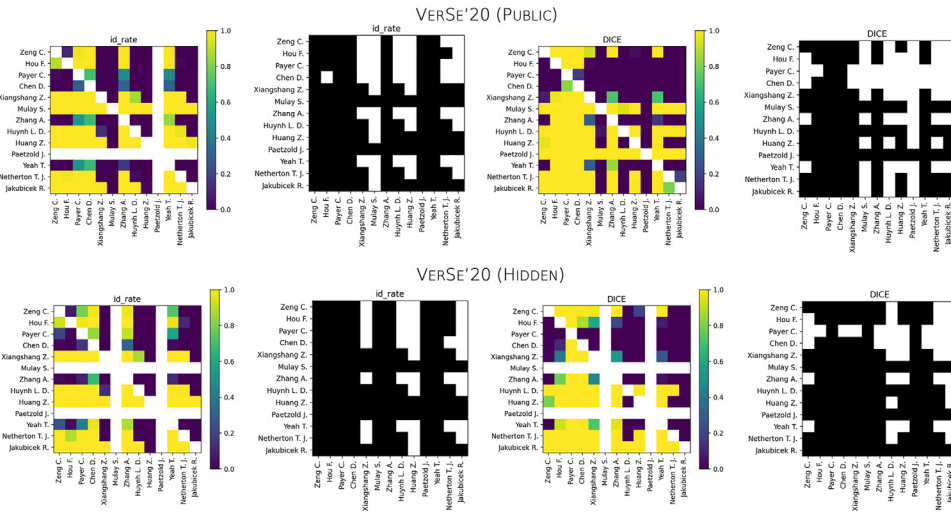


Fig. A.12. VERSE'20 points: Illustrating the p -value matrices and their binarised versions for every metric used.

at a resolution of 1mm with a standard deviation, $\sigma = 8$, and then downsampled to a resolution of 4mm. Additionally, spatial squeeze and channel excite blocks (SSCE) are employed to increase the network's performance-to-parameters ratio. Specifically, the probability of each voxel being a *spine voxel* or a *non-spine* one is predicted by optimising a combination of ℓ_2 and binary cross-entropy losses

as shown:

$$\mathcal{L}_{\text{detect}} = \|\mathbf{y}_d - \hat{\mathbf{y}}_d\|_2 - H(\sigma(\mathbf{y}_d), \sigma(\hat{\mathbf{y}}_d)) \quad (\text{B.1})$$

where \mathbf{y}_d is constructed by concatenating the Gaussian location map with a background channel obtained by subtracting the foreground from 1, $\hat{\mathbf{y}}_d$ denotes the prediction of whose foreground

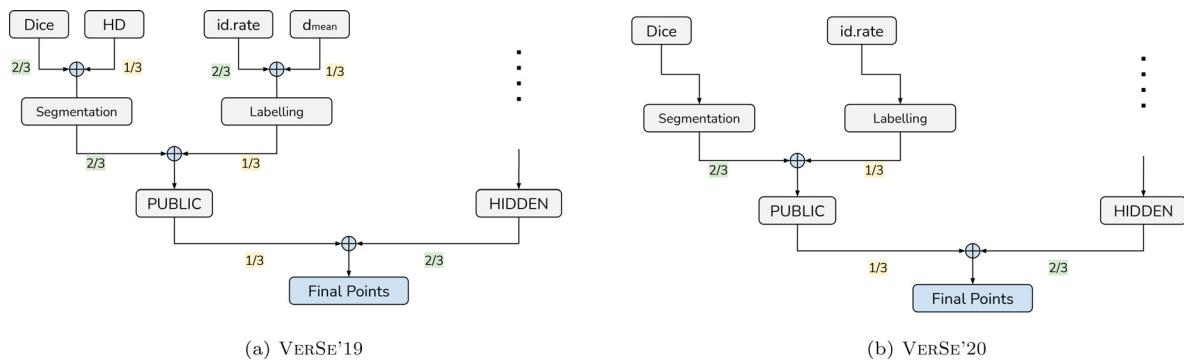


Fig. A.13. Protocol for obtaining the final ranking: Flow diagram of the weights assigned to each stage of the VERSE evaluation, in order to obtain the final point count.

channel represents the desired location map, and $\sigma(\cdot)$ and $H(\cdot)$ denote the softmax and cross-entropy functions.

B3. Stage 2: Vertebrae labelling

To label the vertebrae, we adapt and improve the Btrfly net (Sekuboyina et al., 2018; 2020) that works on two-dimensional sagittal and coronal MIP. By virtue of the spine's extent obtained from the previous component, MIPs can now be extracted from a region focused on the spine, thus eliminating occlusions from ribs and pelvic bones. Cropping the scans to the spine region also makes the input to the labelling stage more uniform, thus improving the training stability. The labelling module works at 2mm isotropic resolution and is trained by optimizing the loss function that is a combination of the sagittal and coronal components, $\mathcal{L}_{\text{label}} = \mathcal{L}_{\text{label}}^{\text{sag}} + \mathcal{L}_{\text{label}}^{\text{cor}}$, where the loss of each view is given by:

$$\mathcal{L}_{\text{label}}^{\text{sag}} = \|\mathbf{y}_l^{\text{sag}} - \tilde{\mathbf{y}}_l^{\text{sag}}\|_2 + \omega H(\sigma(\mathbf{y}_l^{\text{sag}}), \sigma(\tilde{\mathbf{y}}_l^{\text{sag}})), \quad (\text{B.2})$$

where $\tilde{\mathbf{y}}_l^{\text{sag}}$ is the prediction of the net's sagittal-arm of the Btrfly net and ω denotes the median frequency weight map giving a higher weight to the loss originating from less frequent vertebral classes.

B4. Stage 3: Vertebral segmentation

Once the vertebrae are labelled, their segmentation is posed as a binary segmentation problem. This is done by extracting a patch around each vertebral centroid predicted in the earlier stage and segmenting the vertebra of interest. An architecture based on the U-Net working at a resolution of 1mm is employed for this task. Additionally, SSCE blocks are incorporated after every convolution and upconvolution block. Importantly, as there will be more than one vertebra within a patch, a vertebra-of-interest (VOI) arm is used to point the segmentation network to delineate the vertebra of interest. The VOI arm is an encoder parallel to the image encoder as shown in Fig. B.14, processing a 3D Gaussian heatmap centred at the vertebral location predicted by the labelling stage. The feature maps of the VOI arm are concatenated to those of the image encoder at every resolution. The segmentation network is trained using a standard binary cross-entropy as a loss.

B5. Inference & interaction

Simplifying the flow of control throughout the pipeline, Algo. 1 describes the inference routine given a spine CT scan and various points where medical experts can interact with the results, thus improving its overall performance.

Appendix C. Participating Algorithms

Amiranashvili T. et al.: Combining Template Matching with CNNs for Vertebra Segmentation and Identification

A multi-stage approach is adopted to label and segment the vertebrae as illustrated in Fig. C.15: 1. Multi-label segmentation with arbitrary, but separate labels for each vertebra based on local regions of interest in the image. 2. Unique label-assignment to segmented vertebral masks based on shape, while globally regularising over the entire CT field of view. 3. Derive landmark positions from the multi-label segmentation masks by applying a shape-based approach.

Multi-label Segmentation. This stage includes creating a first, rough binary segmentation of the overall spine followed by localising regions of interests around each vertebra and performing voxel-level, refined segmentation of each vertebra. Binary segmentation separating the spine from the background is achieved through a U-Net employed on 2D sagittal slices. For each slice, neighboring slices are included as additional channels in the input to provide a larger context. The network is trained on fixed-size, random crops from original slices. Following this, the number of vertebra and their rough positions are computed based on the binary segmentation by combining shape-based fitting via generalised Hough transform (GHT) (Seim et al., 2008) with a CNN-based heat-map regression for localising vertebra in the spinal column. Put to use in the fitting procedure were manually generated GHT templates of the lumbar (L1-L5), lower thoracic (T10-T12), mid-thoracic (T5-T9), upper-thoracic (T1-T4), lower-to-mid cervical (C3-C5), and upper-cervical (C2-C1) spine. The Butterfly network (Li et al., 2018) was trained on mean and maximum intensity projections in anterior-posterior and lateral directions of the CTs. Finally, multi-label segmentation is performed based on the rough locations from the previous step by deriving a region of interest for each visible vertebra. Individual vertebrae are then segmented via a U-Net based on 2D sagittal slices cropped to the corresponding regions of interests while including neighboring slices as additional input channels. The segmentation masks resulting from the cropped images are then combined into a multi-label segmentation mask.

Vertebra Identification. Vertebra identification is performed based on shape through template fitting along with explicit global regularisation over the whole visible spine. For each vertebra, shape templates are fitted non-rigidly to the given labels via the iterative closest points (ICP) algorithm using the six templates introduced above. This results in a table containing a fitting score for each template and each detected vertebra. Then, optimisation for a set of unique vertebra types is performed such that the combined score from the table is maximised while maintaining the consistent ordering of vertebra (e.g. L4 must follow L5). The multi-

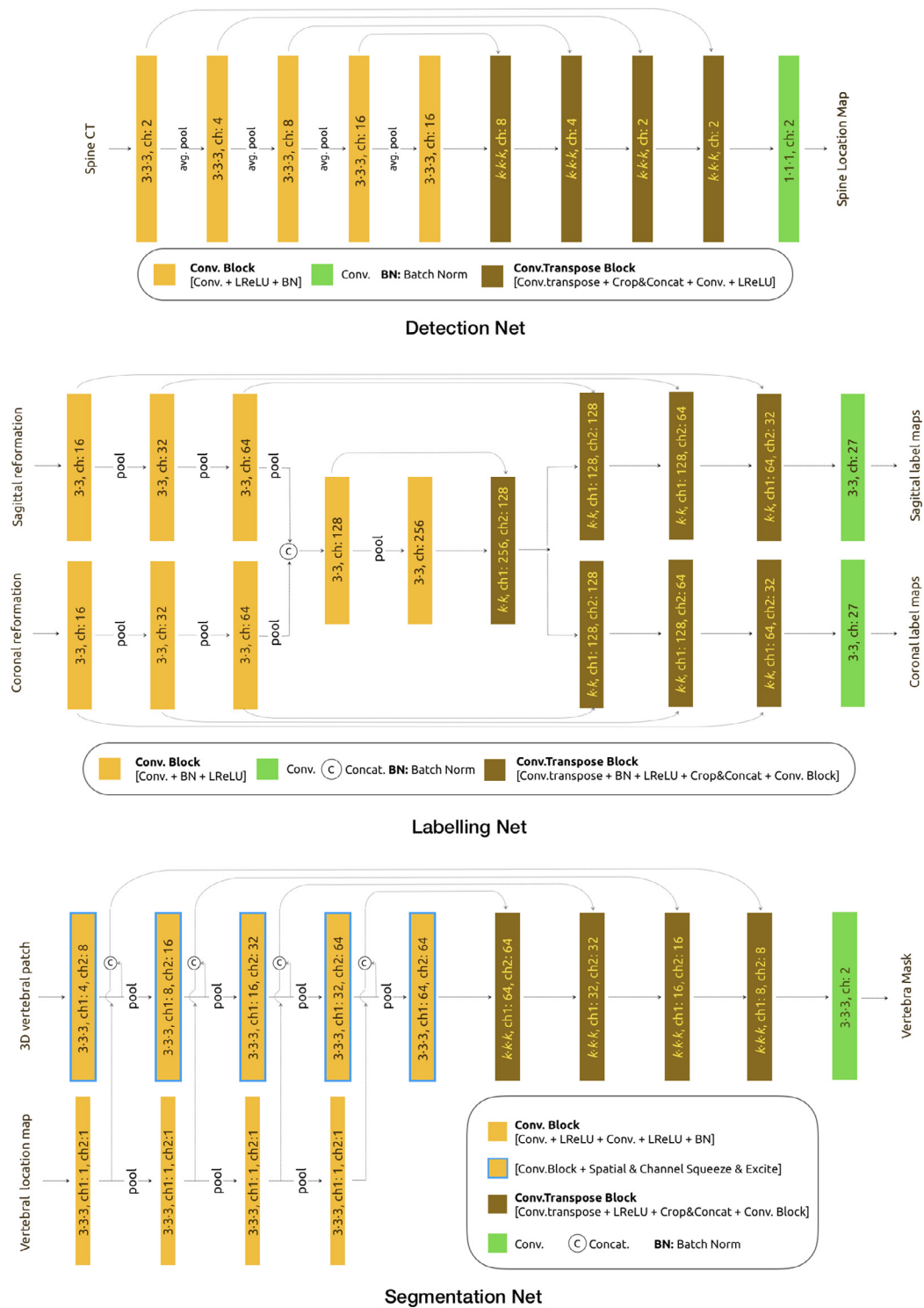


Fig. B.14. Architectures: Detailed network architectures of the three stages in *Anduin*: the spine detection, vertebrae labelling, and the vertebra segmentation stages. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

label segmentation of the previous stage is then re-labelled according to the determined ordering, resulting in a segmentation with uniquely identified labels for each vertebra.

Landmark Extraction. After segmentation and identification, the positions of the landmarks are identified by re-fitting a template of the body of each vertebra to the unique labels followed by extracting the template’s centre point which forms the landmark.

Angermann C. et al.: A Projection-based 2.5D U-net Architecture for VERSE’19. (Angermann et al., 2019)

For the task of a fully-automated technique for volumetric spine segmentation, a combination of a 2D slice-based approach and a projections-based approach is proposed with two tasks: 1. 3D spine segmentation with one output channel denoting the probability of a voxel belonging to a vertebra, followed by assignment of

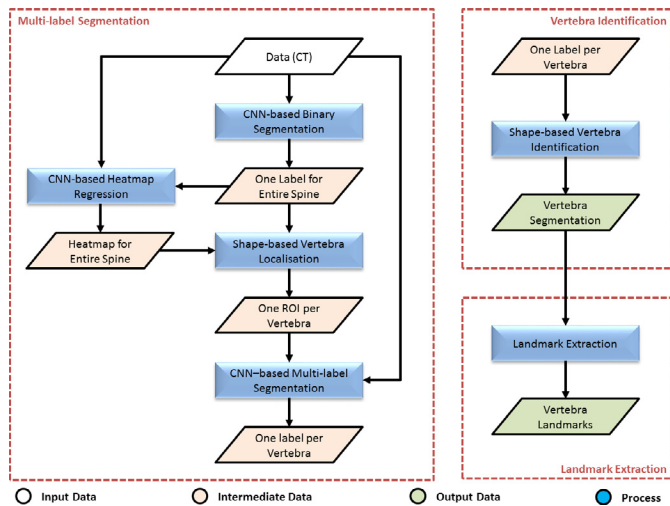


Fig. C.15. Multiple stages involved in the algorithm proposed by Amiranashvili T.

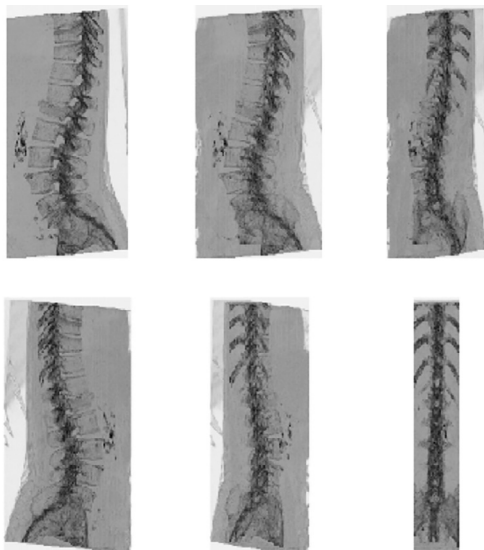


Fig. C.16. Maximum intensity projections of a 3D spine scan with directions $\{k \times 30 \text{ degrees} | k = 0, \dots, 5\}$.

a label from C1 to L6. 2. Using the multi-label segmentation mask, weighted centroid computation for each label for the task of vertebra labelling. Please refer to (Angermann et al., 2019) for details on the 3D segmentation procedure.

Vertebra Segmentation. This is a two-step approach working with images of size $224 \times 224 \times 224$, obtained by zooming the array such that the longest axis is size 224 and padding the other axes with zeros. In the first step, whose output is a one channel segmentation mask (vertebra as foreground), a 2.5D U-net (Angermann et al., 2019) and two 2D U-net are employed. The former network takes the 3D array as input and generates 2D projections containing full 3D information. Here the MIPs are employed (cf. Fig C.16). These 2D projections are propagated through a 2D U-net and lifted back to a volume using a trainable reconstruction algorithm (cf. Eq 3.1, (Angermann et al., 2019)). Due to the non-convex nature of vertebrae, this segmentation is combined with that of a 2D slice-based U-net in the probability space. In the second step, the binary segmentation mask is assigned multiple labels. For this, A 2D U-Net working on six MIPs per scan is employed. Each of the MIPs is obtained at an angle in $\{0^\circ, 10^\circ, 80^\circ, 90^\circ, 100^\circ, 170^\circ\}$, as in Fig. C.16. As output, six labelled

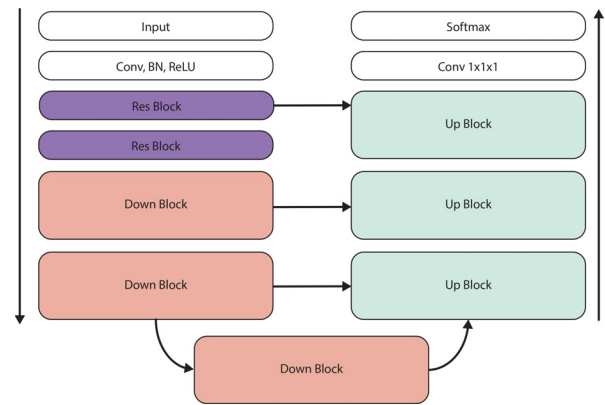


Fig. C.17. The residual U-Net employed for segmentation in Brown K's approach.

MIP segmentation masks are obtained. From these, the 3D labelled mask is obtained by back-projection, wherein each 2D MIP mask is multiplied by a rotated 3D binary segmentation from the previous step, rotated according to the angle corresponding to the MIP mask in question.

Vertebra Labelling. Since the vertebrae are already labelled in the segmentation stage, the vertebral centroids are obtained by just weighing the edges of the vertebra and computing the centroid. The edge-weight is set empirically and is same across the vertebrae.

Brown K. et al.: Spine Segmentation with Registration

Segmentation of the vertebrae is performed by extracting a bounding box around each vertebra and segmenting this box with a residual U-net. The bounding box around the vertebra is identified via a regressed set of canonical landmarks. Each vertebra is then registered to a common 'atlas' space via these landmarks. For segmentation, the employed residual U-net works with inputs of size $64 \times 64 \times 64$ voxels with a depth of five blocks (cf. Fig. C.17).

Objective Function. A network is trained to minimise a combination of Dice coefficient (L_D) and a weighted false-positive/false-negative loss (L_{FPFN}), described as: $L = L_D + \alpha L_{FPFN}$ ($\alpha = 0.5$ in this work). Specifically, the dice coefficient measures the degree of overlap between two sets. For two binary sets ground truth (G) and predicted class membership (G) with (N) elements each, the dice coefficient can be written as

$$D = \frac{2 \sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i}, \quad (C.1)$$

where each p_i and g_i are binary labels. In this case, p_i is set to $\{0, 1\}$ from the softmax layer representing the probability that the i^{th} voxel is in the foreground class. Each g_i is obtained from a one-hot encoding of the ground-truth-labelled volume of tissue class. Additionally, the weighted false-positive/false-negative loss term is included to provide smoother convergence. It is defined as:

$$L_{FPFN} = \sum_{i \in I} w_i p_i (1 - g_i) + \sum_{i \in I} w_i (1 - p_i) g_i, \quad (C.2)$$

where the weight, $w_i = \gamma_e \exp(-d_i^2/\sigma) + \gamma_c f_i$, with d_i being the euclidean distance to the nearest class boundary and f_i the frequency of the ground truth class at voxel i . In this work, σ is chosen to be 10 voxels, and the parameters γ_e and γ_c are set to 5 and 2, respectively.

Chen M.: An Automatic Multi-stage System for Vertebra Segmentation and Labelling

A three-stage strategy is applied to solve the task of vertebral segmentation and labelling. The first two stages are based on a U-Net architecture for multi-label segmentation. Utilising the predicted segmentation mask, the third stage employs an RCNN-based architecture (Girshick et al., 2014; Girshick, 2015) to label the vertebrae.

Segmentation (Stages 1 & 2). The first stage consists of a 3D U-Net working on randomly extracted patches of size $224 \times 160 \times 128$. The network is trained to predict 25 labels, ignoring the rare L6 label. It is observed that the segmentation Stage 1 performs well in regions close to C1 and L5. However, in the other regions, the vertebral labels are mixed with each other due to a similarity in their shapes. To resolve this problem, a second *refinement network* is introduced with an architecture similar to the first stage but with a major difference in the training regime. For this, patches are extracted covering the spine in the middle and extending 1.5 times in the *slice* direction. These patches are padded to $128 \times 128 \times 128$ with zeroes if necessary. The network is trained to predict a binary label only for the mid-vertebra. The combination is trained as follows: All the labelled Stage 1 masks are combined into a binary mask, indicating the foreground. Each of these masks (corresponding to each vertebral label) is used to generate a patch for Stage 2. This prediction is believed to be accurate at instance level and filled back into the binary foreground. If the foreground is not filled sufficiently, new patches will be selected from the not-filled regions for Stage 2 recursively till convergence. Because the well-segmented instances in Stage 1 and Stage 2 mostly overlap, it is operable to assign labels based on both the stages by comparing the Dice of the pairs. With the constraint on the label continuity of neighboring spines, this process can be performed using the matching algorithm presented in Fig. C.18.

Labelling. An RCNN-based architecture with a 3D ResNet-50 is used as the backbone for the vertebra labelling task. RoI pooling is performed on the features of the feature map at stride 4 to regress the deviation of the vertebra centre to the RoI box's centre in the coordinate space of the box. This network works with inputs of size $160 \times 192 \times 224$. In the training phase, boxes are generated from the segmentation ground truth such that more positive samples are generated. During inference, the predicted segmentation mask is utilised.

Dong Y. et al.: *Vertebra Labeling and Segmentation in 3D CT using Deep Neural Networks* (Yu et al., 2020)

A U-shaped deep network is used for generating the vertebral segmentation masks and labels in the form of a model ensemble followed by a post-processing module.

The problem is formulated as a 26-class segmentation task given 3D CT as input. The class information from prediction is able to provide labels (cervical C1 ~ C7, thoracic T1 ~ T12, lumbar L1 ~ L6) for different vertebrae. For vertebra localisation, the centroids of vertebrae are determined as the mass centres of segmentation masks.

We have adopted a U-shaped neural network for vertebral segmentation following the fashion of the state-of-the-art network for 3D medical image segmentation. The network architecture is nearly symmetric with an encoder and a decoder. After achieving the segmentation results, the vertebrae centroids are computed based on the mass centres of binary labels for each individual vertebra. To further help determine the vertebral body centre, several iterations of morphological erosion are conducted to remove the vertebral 'wings'. The final prediction is from the ensemble of the five models.

Algorithm 1 Update label for stage-2 vertebrae set

Input: Stage-2 vertebrae set V_n ($V_n = 1, 2, \dots, k$) and the stage-1 vertebrae set V_r (size= $m, 1 \leq \max(V_r) \leq 26$)

Output: Updated vertebrae label set

```

1: if Stage-1 vertebrae set contain label 22 or 23 and  $m \leq 12$  then
2:   for instance  $i \in V_n, i = k, k - 1$  do
3:     for vertebra  $v_j \in V_r, v_j \geq i$  do
4:       Calculating and recording dice index for instance  $i$  with vertebra  $v_j$ 
5:     end for
6:     Find the Maximum of record dice and the corresponding vertebra  $v_b$ 
7:     if maximum of record dice  $\geq 0.8$  then
8:       if  $i = k$  then
9:         update label for stage-2 vertebrae set from  $v_b - k + 1$  to  $v_b$ 
10:      else
11:        update label for stage-2 vertebrae set from  $v_b - k + 2$  to  $v_b + 1$ 
12:      end if
13:      break
14:    end if
15:  end for
16: else
17:   for instance  $i \in V_n, i = 2, 3, 4$  do
18:     for vertebra  $v_j \in V_r, i \leq v_j \leq 25 - k + 1$  do
19:       Calculating and recording dice index for instance  $i$  with vertebra  $v_j$ 
20:     end for
21:     Find the Maximum of record dice and the corresponding vertebra  $v_b$ 
22:     if maximum of record dice  $\geq 0.8$  then
23:       if  $i = 2$  then
24:         update label for stage-2 vertebrae set from  $v_b - 1$  to  $v_b + k$ 
25:       else if  $i = 3$  then
26:         update label for stage-2 vertebrae set from  $v_b - 2$  to  $v_b + k + 1$ 
27:       else
28:         update label for stage-2 vertebrae set from  $v_b - 3$  to  $v_b + k + 2$ 
29:       end if
30:       break
31:     end if
32:   end for
33: end if

```

Fig. C.18. Procedure for label correction after Stage 2 of Chen M.'s approach.

Hu Y. et al.: *Large Scale Vertebrae Segmentation Using nnU-Net*

The tasks at hand are posed as an application of the nnU-Net (Isensee et al., 2019), a framework that automatically adapts the hyper-parameters to any given dataset.

Generally, nnU-Net consists of three U-Net models (2D, 3D, and a cascaded 3D network) working on the images patch-wise. It automatically sets the training hyper-parameters such as the batch size, patch size, pooling operations etc. while keeping the GPU budget within a certain limit. If the selected patch size covers less than 25% of the voxels in case, the 3D-Net cascade is additionally configured and trained on a downsampled version of the training data. Specific to VERSE'19, a sum of cross-entropy loss and Dice loss are used the training objective, minimised using the Adam optimiser. An initial rate of 3×10^{-4} and ℓ_2 weight decay of 3×10^{-5} . The learning rate is dropped by a factor of 0.2 whenever the exponential moving average of the training loss does not improve within the last 30 epochs. Training is stopped when the learning rate drops below 10^{-6} or 1000 epochs are exceeded. The data is augmented using elastic deformations, random scaling, random rotations, and gamma augmentation. Note that in Phase 1, the nnU-Net ensemble did not include all its components. Included are a 3D U-Net operating at full resolution, a 3D U-Net at low resolution (as part of the cascade 3D), and a 2D U-Net.

Jiang T. et al.: *SpineAnalyst: A Unified Method for Spine Identification and Segmentation*

In contrast to most approaches that treat identification and segmentation as two separate steps, this work efficiently solves them simultaneously with a keypoint based instance segmentation framework applying anchor-free instance segmentation networks in a 3D setting. To the best of the participant's knowledge, this is a

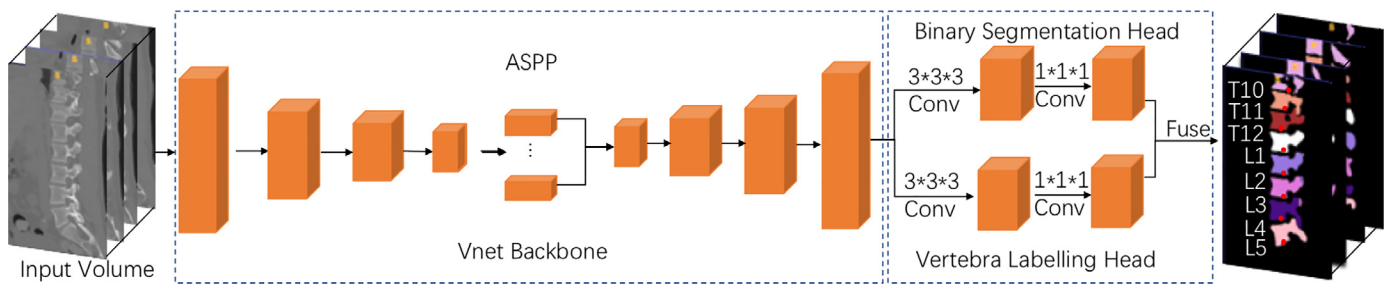


Fig. C.19. An overview of SpineAnalyst network, a contribution of Jiang T..

first. The proposed network adopts the encoder-decoder paradigm with two prediction heads attached to the shared decoder, as described in Fig. C.19. The “binary segmentation head” distinguishes spine pixels resulting in a binary semantic map. The “vertebra labeling head” detects and labels all the vertebrae landmarks, while also predicting a vector field that associates vertebral pixels with their vertebrae centres. The predictions of two heads are fused together to produce the final instance segmentation results

Encoder & Decoder. A V-Net is used as the backbone with the encoder containing four cascaded blocks. Following this, the Atrous Spatial Pyramid Pooling (ASPP) method is applied to further increase the receptive field and capture multi-scale information effectively. In the decoder, the concatenated features of ASPP are passed through four cascaded up-sampling blocks recovering the original volume resolution.

Binary Segmentation Head. A binary semantic segmentation head is trained to detect the spine as the foreground pixels. These pixels will further be assigned with vertebral labels in the subsequent fusion processing.

Vertebra Labeling Head. This component performs two tasks: 1. Detect and label landmarks: For the former, the heatmap channels predict the probability that a pixel belongs to a vertebra centre. Pixels corresponding to high confidence are reserved as vertebral landmarks. Due to the similarity of the adjacent vertebrae, it is challenging to directly identify individual vertebrae. Instead, the reference vertebrae with obvious anatomical features, such as C2, L5 and C7, T12, are first identified. Other vertebrae labels are then inferred from the reference vertebrae. Following this, 2. a vector-field is predicted with each channel denoting the offsets relative to the corresponding vertebra centre. Each pixel is then labelled with the closest vertebra centre according to the long offset.

Fusion Process. The final instance segmentation is obtained from binary semantic segmentation as follows: Each pixel within the semantic mask acquires its label from the centre point closest to its predicted centres, which is computed by pixel coordinates plus the vector field.

Kirszenberg A. et al.:

A multi-stage approach is proposed involving a pseudo-3D U-Net architecture for segmentation and a template matching approach enabled by morphological operation.

Segmentation. Three different U-Net models are trained in a “pseudo-3D” segmentation technique wherein, the 3D input is sliced into 3-voxel wide slices along the three axes. Prior to this, patches of size $80 \times 128 \times 128$ are extracted from the scan, resulting in sagittal, coronal, and axial slices of shapes $3 \times 123 \times 128$, $80 \times 3 \times 128$, and $80 \times 128 \times 3$, respectively. This step performs a binary segmentation of “spine vs. background”. The predicted masks of the three models are combined using majority voting and

passed through a filtering operation for removal of stray segmentation and hole-filling (cf. Fig. C.20a).

Labelling. This task is attempted as a combination of morphological operations and template matching, implemented as follows: 1. The predicted binary segmentation mask is blurred using a Gaussian kernel and skeletonised to obtain a skeleton of the vertebral column. Further clean-up is obtained by choosing the path connecting the voxels between two end-points using Dijkstra’s algorithm. 2. The skeleton is then discretised into 1mm distant points which are used as anchors for template matching. These templates were generated from the training data at a vertebra level by centring each vertebra at the centroid and averaging over a certain number of rotations as shown in Fig. C.20b. For template matching, the five best vertebrae, point candidates are chosen and for every point its previous and next vertebrae are matched to the points before and after, respectively. Once no vertebrae can be matched, scores for each vertebrae are summed from each of the five vertebral columns and the one with the highest score is selected. Following this, each voxel of the column is labelled after the template with the highest score.

Wang X. et al.: Improved Btrfly Net and a residual U-Net for VERSE’19

Improved versions of Btrfly Net (Sekuboyina et al., 2018) and the U-Net (Ronneberger et al., 2015) are employed to address the tasks of labelling and segmentation, respectively. Of interest is the task-oriented pre- and post-processing employed in each task.

Pre-processing. A Single Shot MultiBox Detector (SSD) is implemented to localise the vertebrae in the sagittal and coronal projections and its predictions are used to crop the 3D scans. This is followed by re-sampling the crops to a 1mm resolution and padding the projections to 610×610 pixels.

Labelling. The Btrfly Net is employed for this task with a major difference in the reconstruction of 3D coordinates from its 2D heatmap predictions. However, unlike obtaining the 3D coordinates from the outer product of the 2D channelled heat-maps followed by an *argmax*, the authors propose an improved scheme resulting in a 4% improvement of the identification rate. Specifically, 2D coordinates of the vertebra are obtained from the individual projections, denoted by (x, z_s) from the sagittal and (y, z_c) from the coronal heat maps. Notice the two variants of the z -coordinate. The final z -coordinate is then calculated as the weighted average of z_s and z_c with the maximum values of their corresponding heat maps as weights. Additionally, the missing predictions are filled in with interpolation.

Segmentation. Since the vertebral centroids are now identified, the segmentation is tasked to segment one vertebra given its centroid position. For this, a 3D U-Net with residual blocks is chosen (Fig. C.21). The network is trained with Dice loss and works with patches of size $96 \times 96 \times 96$ centred at the vertebral centroid in question. Once segmented, the vertebra is labelled according to its centroid’s label and assigned back to the full scan. In case of a con-

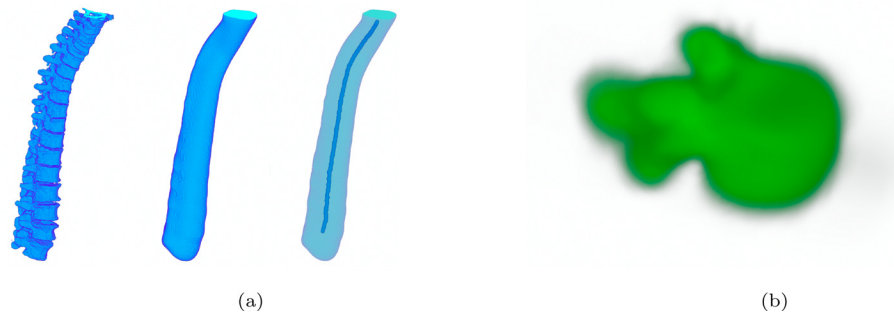


Fig. C.20. Team Kirszenberg A.'s contribution involving (a) detection of the spline passing through the vertebral column and (b) a sample template for L4 use for vertebra identification.

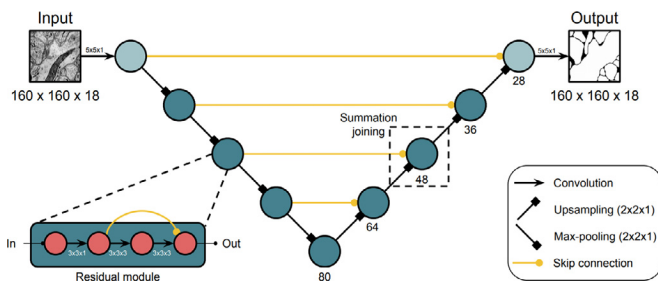


Fig. C.21. Architecture of residual U-net employed by team Wang X. for the segmentation task.

flict, i.e.: if a voxel labelled as i is again labelled as j , the label with a higher logit is chosen.

Hou et al.: Fully Automatic Localisation and Segmentation of Vertebrae Based on Cascaded U-Nets

The authors propose a multi-stage pipeline for vertebral localisation and segmentation based on a general U-net architecture. Firstly, the centre-line of the spine is inferred, and then the spine region is cropped to be fed as the input of the second stage. Accordingly, the second neural network predicts the centre coordinates and classes of all vertebrae. In the last stage, the segmentation network performs a binary segmentation of each of the cropped vertebrae. The full pipeline is illustrated in Fig. C.22

Spine Localisation. In the first stage, the authors use a variant of the U-Net (Ronneberger et al., 2015) to predict heat-maps that cover the whole spine. They set the filters of each convolutional layer to 64, which can significantly improve training speed while ensuring performance. The authors utilise the general ℓ_2 -loss to minimise the difference between the target and predicted heat-maps. As a pre-processing step, the CT images are sub-sampled to a uniform voxel spacing of 8mm, and then a patch size of $64 \times 64 \times 128$ is fed into the network. The predicted coordinates of the centre of the spine help are used to crop the spine region as the input of the second stage.

Vertebrae Localisation. The authors deploy the general U-Net (Ronneberger et al., 2015) as a baseline. Both encoder and the decoder use five levels consisting of two convolution layers with a leaky-ReLU activation function. Due to the specific shape and fixed relative position of vertebrae, for most cases, the labels of the vertebrae are a continuous sequence despite their coordinates. It is important to localise and identify the first and the last vertebrae. The authors use a weighted ℓ_2 -loss function to emphasise the con-

tribution of the first and the last vertebrae in the loss. Similarly to the first stage, the CT images are re-sampled to uniform voxel spacing of 2mm, and then a patch size of $96 \times 96 \times 128$ is fed into the network.

Vertebrae Segmentation. In this stage, the predicted coordinates of each vertebra are used to crop the individual vertebrae region. Similar to the localisation stage, the U-Net is used and the CT volumes are re-sampled to a uniform voxel spacing of 1mm, the segmentation network with a patch size of $128 \times 128 \times 96$ produces the individual predictions of each vertebra, and finally, the multi-label segmentation results are obtained by merging all binary segmentation results.

Postprocessing. Due to the partial vertebrae often in the top or bottom of volume, which has a bad influence on detecting the position of the first or last vertebrae, in this work, the landmark is abandoned if its distance from the top or the bottom of the volume is less than a threshold.

Huang et al.: A²Unet: Attention and Aggregation UNet for Vertebrae Localisation and Segmentation

The authors formulate both tasks as a pixel-level prediction problem. Specifically, the landmark detection problem (task 1) is converted into a heat-map prediction format and the vertebrae segmentation problem (task 2) is converted into a multi-class semantic map prediction scheme. Both tasks generate full-scale outputs that enabled the authors to utilise a U-net architecture (Ronneberger et al., 2015) to extract the features. In this work, the authors develop a new variation of 3D U-net, in which an attention and aggregation mechanisms are introduced to enhance the feature representation in both tasks. This new variant is called the A²UNet.

Attention and Aggregation UNet (A²UNet). The proposed A²UNet, which is shown in Fig. C.23, adopts the original U-Net structure that consists of a contracting path to downsample the inputs for global representation, and an expanding path to upsample the feature for detailed prediction. Several skip connections link the contracting and expanding paths that directly transfer the information from the shallow to deep layers. However, features from different convolution stages contain information from different semantic levels. In this work, the authors embed the efficient feature aggregation (FA) module shown in Fig. C.24, into the U-Net structure for channel-wise attention based on the Squeeze-and-Excitation (SE) block (Hu et al., 2019). It receives the two feature maps where one is from the contracting path, and the other is from the expanding path. The features are firstly sent to the average pooling process for global representation. Then, two fully connected layers are used to investigate the importance (weights) of different feature

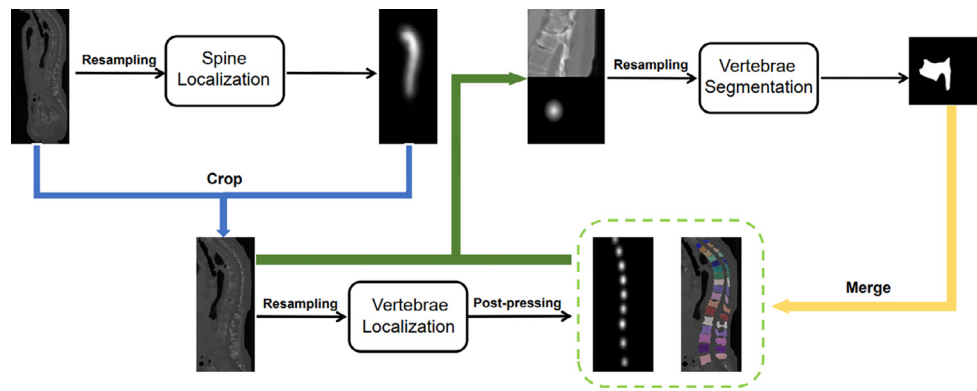


Fig. C.22. Team Hou et al.'s proposed pipeline.

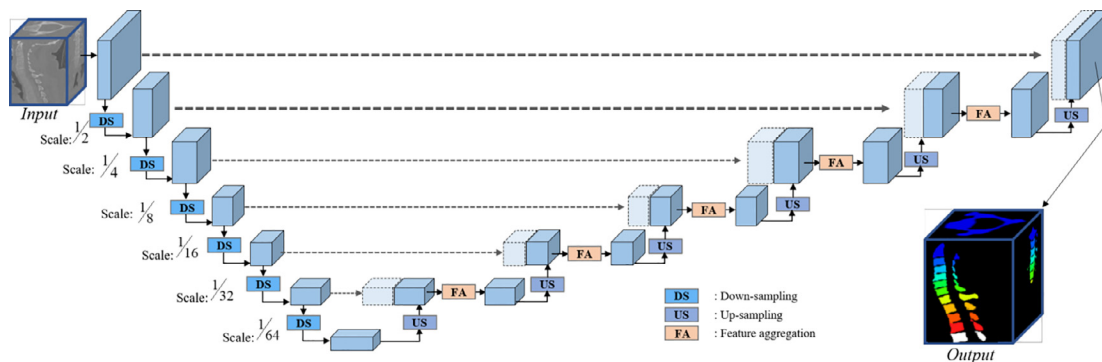


Fig. C.23. A²Unet's Architecture.

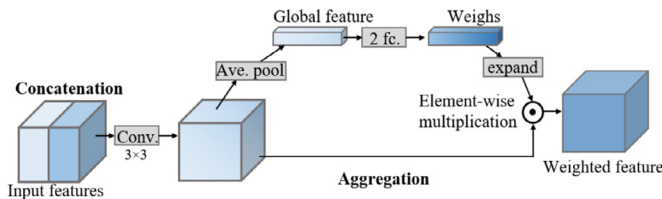


Fig. C.24. Team Huang et al.'s Feature Aggregation block.

channels. By multiplying the weights to corresponding channels, the key features can be focused that will be used for the following process.

Heads for Vertebra Localisation and Segmentation. The authors develop two sub-networks, or heads, to decode the backbone output into the feature format for each task. For the localisation task, a convolution layer is applied to generate a 26-channel output, each channel corresponds to one of 26 classes of the vertebra. Each channel is actually a heat map where the location information of a specific vertebra is encoded. To reason the vertebra location, the coordinates candidates are selected where the corresponding score in the heat-map is above 0:35. The final vertebra coordinates are determined by adopting the non-maximum suppression (NMS) algorithm towards those candidates in an adjacent vertebra region which has a distance between 12.5mm and 40mm.

For the segmentation task, a convolutional layer is deployed to generate a single semantic map, each pixel contains 27 categorical value, indicating one of 26 anatomical classes or the background. The segmentation model is trained with Dice loss and CE loss. Since every voxel is classified only considering the channel score after obtaining the segmentation mask, outlier voxels that are not connected with the largest component will be removed.

Huỳnh et al.: 3D Mask Retinanet for Vertebrae Instance Segmentation

The authors propose a single model that performs both sub-tasks. A two-stage model is adopted inspired by Mask R-CNN (He et al., 2018). Mask R-CNN is a two-staged model, in which the first stage localises RoI while two sub-nets on the second stage classify and segment a subset of these RoIs. Since the Mask R-CNN is a heavy model, an extended version or Mask R-CNN for 3-D images will require significant memory, and as a result, it limits the number of RoIs that could be passed to the second stage. This problem makes the model more sensitive to class imbalance. For that reason, the authors propose a new two-stage model. They replace the first stage of Mask R-CNN with the Retinanet (Lin et al., 2018). With this modification, the first stage is now responsible for both RoIs" localisation and classification. The first stage is more robust to class-imbalance than the original Mask R-CNN thanks to Focal Loss. This allows the authors to use a small, fully convolutional network on the second stage to performs the mask regression. Since only RoIs that contain objects will be passed through the second stage, training the model requires less memory. This model is called the Mask RetinaNet. Due to memory limitation, the authors are forced to train with a small batch size and they use Group Normalisation (Wu and He, 2018) instead of Batch Normalisation in their network. The architecture of Mask RetinaNet is illustrated in Fig C.25

The detector stage. The authors adapt RetinaNet for 3D cases. Their version will also predict the object's centroid in addition to the axis-aligned bounding box (AABB). The backbone is constructed with a 3D version of the Resnet50 and Feature Pyramid Network (Lin et al., 2018). For this dataset, the authors only use pyramid levels 3 to 5. They avoid level 2 because anchors defined on it are unnecessarily dense for this dataset, while an-

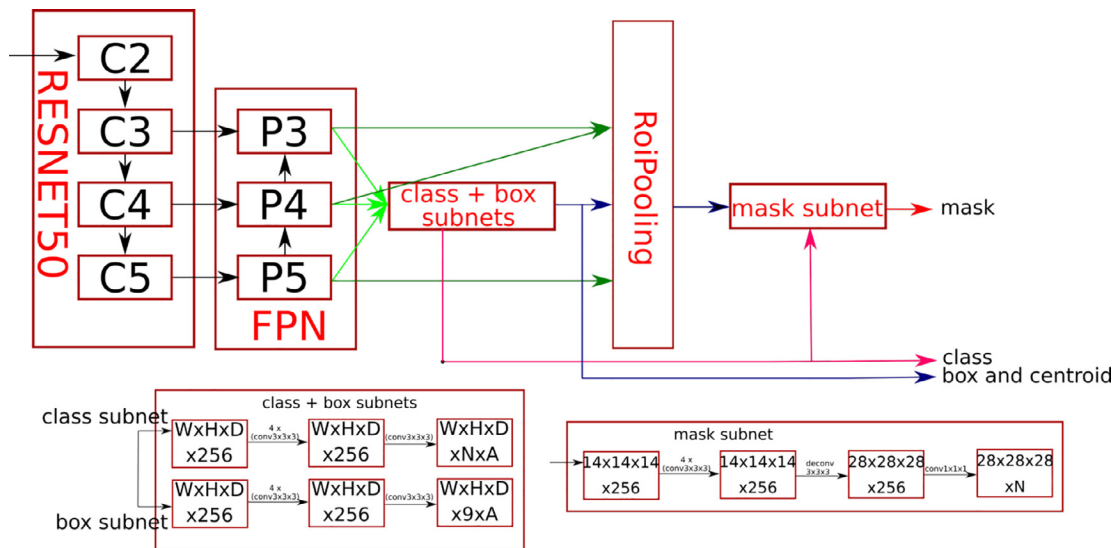


Fig. C.25. Mask RetinaNet's architecture, as employed by *Huynh et al.*

chors defined on levels greater than 5 are too sparse to distinguish nearby vertebrae. At each pyramid level, they use four anchors with two width/height/depth ratios of 1/1/0.625 and 1/0.74/0.42, and a width of 86 and 68 for level 3, 100 and 79 for level 4 and 120 and 94 for level 5. They are chosen by running a K-means clustering on the AABBs of training vertebrae similar to the algorithm described in (Redmon and Farhadi, 2016) to ensure that for each vertebra, they could find at least one anchor so that the Intersection over Union (IoU) with its AABB is higher than 0.6. The classification and regression sub-nets are implemented as described in (Lin et al., 2018). The classification subnet is responsible for the classification of anchors. It predicts a length N one-hot classification vector for each anchor, with N being the number of classes. The regression subnet performs AABB and centroid regression. For each positive anchor, it predicts a length nine regression vector, of which the first 6 encode the AABB (its centre coordinate and size), and the last 3 encode the centroids' position. Instead of predicting these values directly, they adopt the coordinate parameterisations of (Ren et al., 2015) for their case.

The mask regression stage. To perform instance segmentation, the authors attach a second stage to their 3D-RetinaNet to output a binary mask for each RoIs detected by the first stage. This stage is implemented similar to Mask R-CNN: a 3D-ROIAlign layer extracts a fixed-size $w \times h \times d$ feature map from the FPN for each RoI using trilinear interpolation, followed by simple fully convolutional networks. These subnets will produce $D \times 2w \times 2h \times 2d \times N$ with D the number of detections provided by the first stage and N the number of classes.

Jakubicek et al.: Approach for Vertebrae Localisation, Identification and Segmentation

The authors propose a fully automatic multi-stage system as shown in Fig. C.26. Moreover, they provide an auxiliary semi-automatic mode that enables the inspection and possibly correction of automatically detected positions of the inter-vertebral discs (IVD) and their labels before the following segmentation step. Their approach combines modern deep-learning-based algorithms with more classical image and signal processing steps and with segmentation using the intensity vertebra models adaptation.

Pre-processing. The authors first attempt to cut the data from background and "black" artefacts caused by geometrical shearing. The second step is the correction of the random rotation, which is

not presented in the real CT data. For this purpose they provide the CTDeepRot algorithm (Jakubicek et al., 2020), which predicts these rotational angles using a CNN and transforms the data into the standard Head First Supine (HFS) patient position.

Detection of spinal cord centre-line: First, each axial slice of the CT data is classified by a CNN into four categories (slices containing complex C1-2, slices with the main part of the spine from C3 to L6, slices containing the sacrum, and remaining areas feet, head, background). A pre-trained AlexNet (Krizhevsky et al., 2012) CNN is used for this purpose. In the slices containing the main part of the spine, the approximate position of a spinal canal is found (Simonyan and Zisserman, 2014) architecture. Each detected centroid of a detected bounding box is taken as a potentially correct centre of the spinal canal in the appropriate axial slice. The whole spinal canal is then traced by the algorithm using the growing inscribed circles, where the detected centroids are taken as starting (seed) points of the tracing. The optimum spine centre-line is then chosen by the population-based optimisation process.

Vertebra localisation and identification. The spine CT data is geometrically transformed into the straightened data according to the spine centre-line curvature. In the straightened data, the centroids of the vertebral bodies are determined by morphological transforms, and the respective intensity profile along the z-axis is taken. This way, the obtained 1D signal is processed by an adaptive IIR (infinite impulse response) filter, which enables detection of the positions of the individual IVDs. Adaptation of the filter is controlled by a statistical model using knowledge about the anatomy of the spine. Finally, each detected IVD is classified into a category of the vertebral type (label) by a combination of a CNN (pre-trained Inception V3 (Guan et al., 2019)) and the dynamic programming optimisation. All used pre-trained CNN architectures were pre-trained on the ImageNet dataset (Russakovsky et al., 2015) and fine-tuned on the authors' database of CT image data.

Vertebra segmentation. The segmentation of the vertebrae is based on four-step vertebra intensity model registration. In the first step the mean model of the individual vertebra is scaled and deployed along the spine in accordance with the detected and labelled IVDs. The second step performs rigid registration of each vertebra, which aligns the model into an optimally precise position in the 3D CT data, followed by improvement via elastic registration of each vertebra. In the third step, the elastic registration is performed on the whole spine model, where the models fits the shapes of the vertebrae. In the last step, the final segmentation

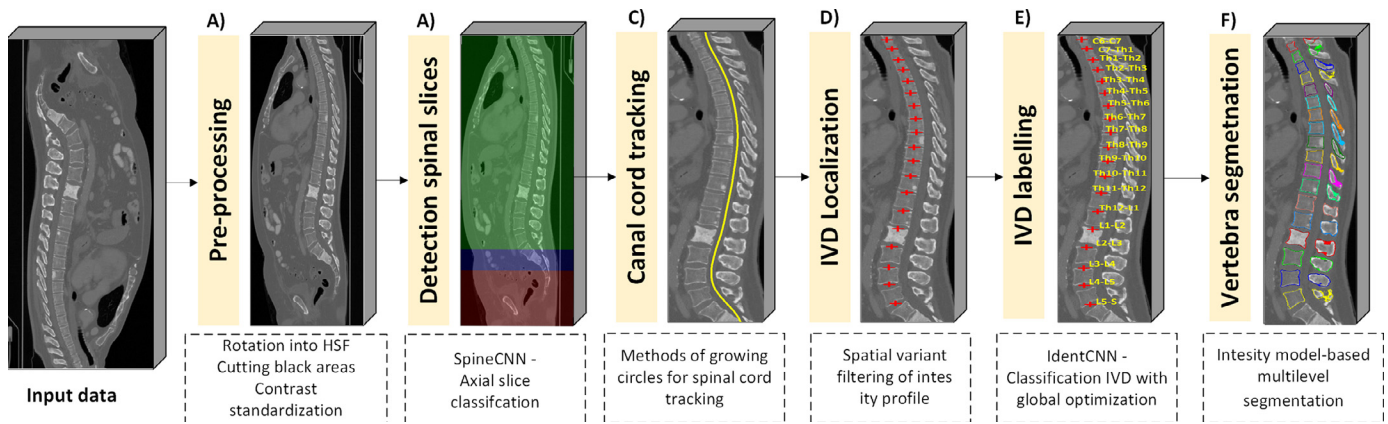


Fig. C.26. An overview of the multi-stage framework proposed by *Jakubicek R.*: Pre-processing, spinal slices detection, spinal canal tracking, inter-vertebral disc (IVD) localisation, IVD labelling, and vertebral segmentation.

contours are slightly refined and smoothed by the graph-cut based algorithm. Elastix v.5.0.0 (Klein et al., 2009; Shamonin et al., 2014) is used as the registration software.

Supriti M. et al.: Vertebrae localisation and Segmentation using Mask-RCNN with Complete-IOU Loss

The authors propose to segment vertebrae using Mask-RCNN trained with Complete-IOU (CIoU) loss. The spine vertebrae segmentation process contains the following pipeline: 3D to 2D conversion, pre-processing, Mask-RCNN feature extraction with Complete IOU loss for geometric factor enhancement (Frosio and Kautz, 2018), and 2D to 3D back conversion.

3D to 2D conversion. Reorientation of the image is done with flips and reordering the image data array so that the axes match the directions indicated in orientation required for spinal vertebrae segmentation. Reoriented images are resampled to get the balance between image smoothness and identify fine image details.

Preprocessing. CT images reconstructed from low-dose acquisitions may be severely degraded with noise and streak artefacts due to quantum noise, or with view-aliasing artefacts due to insufficient angular sampling. To improve CT image quality median filter along with non-local means (NLM) with Statistical Nearest Neighbors (SNN) by Frosio et al. (Frosio and Kautz, 2018) filtering algorithm is applied. Sampling neighbors with the nearest neighbour approach introduces a bias in the denoised patch which improves the CT image quality significantly. Fig. C.27(a) and (c) shows the original slice of a spine CT while (b) and (d) shows the enhanced images.

Segmentation using Mask-RCNN with CIoU loss. Mask R-CNN predicts bounding boxes and corresponding object classes for each of the proposed region obtained using a backbone. Following this, a binary mask classifier generates a mask for every class. Bounding box regression is sometimes inaccurate due to overlapping areas. So a complete IOU (CIoU) loss is added in Mask R-CNN.

A good loss for bounding box regression should consider three important geometric factors, i.e. overlap area, central point distance and aspect ratio. Zheng et al. (Zheng et al., 2020) proposed CIoU loss based on these requirements. The authors use an end-to-end pretrained Mask R-CNN-based detectron with CIoU loss model with Resnet x-152 backbone. An existing open-source implementation³ using Pytorch is chosen.

³ <https://github.com/Zzh-tju/DIoU-pytorch-detectron>.

Once the scan is segmented slice-wise in 2D, the final segmentation is obtained by stacking the predicted masks and reorienting and resampling it back to the original image particulars.

Netherton T. et al.: A Multi-view Localisation and Deeply Supervised Segmentation Framework

The authors propose a framework that combines the use of a set of individual CNNs to accomplish 1) coarse spinal canal segmentation, 2) spine localisation via a multi-view network, and 3) automatic segmentation of individual vertebrae using a deeply supervised approach. A detailed description of steps (1) and (2) of the approach can be found in Netherton et al. (2020). Refer to Fig. C.28 for an overview of the proposed approach.

Algorithm 1: Pseudocode for inference on *Anduin*.

Input: \mathbf{x} , a 3D MDCT spine scan
Output: Vertebral centroids & segmentation masks

~DETECTION

- 1 $\mathbf{x}_d = \text{resample_to_4mm}(\mathbf{x})$
- 2 $\mathbf{y}_d = \text{predict_spine_heatmap}(\mathbf{x}_d)$
- 3 $bb = \text{construct_bounding_box}(\mathbf{y}_d, \text{threshold}=T_d)$
- 4 **Possible interaction:** Alter bb by mouse-drag action.

~LABELLING

- 5 $\mathbf{x}_l = \text{resample_to_2mm}(\mathbf{x})$
- 6 $bb = \text{upsample_bounding_box}(bb, \text{from}=4\text{mm}, \text{to}=2\text{mm})$
- 7 $\mathbf{x}_{sag}, \mathbf{x}_{cor} = \text{get_localised_mips}(\mathbf{x}_l, bb)$
- 8 $\mathbf{y}_{sag}, \mathbf{y}_{cor} = \text{predict_vertebral_heatmaps}(\mathbf{x}_{sag}, \mathbf{x}_{cor})$
- 9 $\mathbf{y}_l = \text{get_outer_product}(\mathbf{y}_{sag}, \mathbf{y}_{cor})$
- 10 $\text{centroids} = \text{heatmap_to_3D_coordinates}(\mathbf{y}_l, \text{threshold}=T_l)$
- 11 **Interaction:** Insert missing vertebrae, delete spurious predictions, drag incorrect predictions.

~SEGMENTATION

- 12 $\mathbf{x}_s = \text{resample_to_1mm}(\mathbf{x}); \text{mask} = \text{np.zeros_like}(\mathbf{x}_s)$
- 13 **for every centroid in centroids do**
- 14 $p = \text{get_3D_vertebral_patch}(\mathbf{x}_s, \text{centroid})$
- 15 $p_{\text{mask}} = \text{binary_segment_vertebra_of_interest}(p)$
- 16 $p_{\text{mask}} = \text{index_of}(\text{mask}, \text{centroid}) * p_{\text{mask}}$
- 17 $\text{mask} = \text{put_vertebrae_in_mask}(p_{\text{mask}})$
- 18 **end**

Data. Data from VERSE 2019 and 2020 was used to train localisation and segmentation CNNs. All images and segmentations were

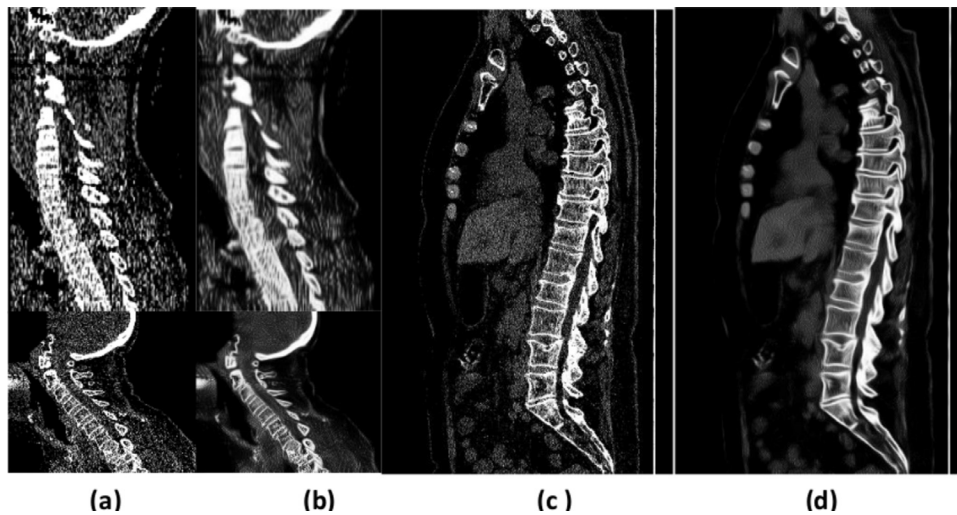


Fig. C.27. Enhancement of CT slices using the filtering algorithm proposed by Frosio and Kautz (2018) employed by Mulay S..

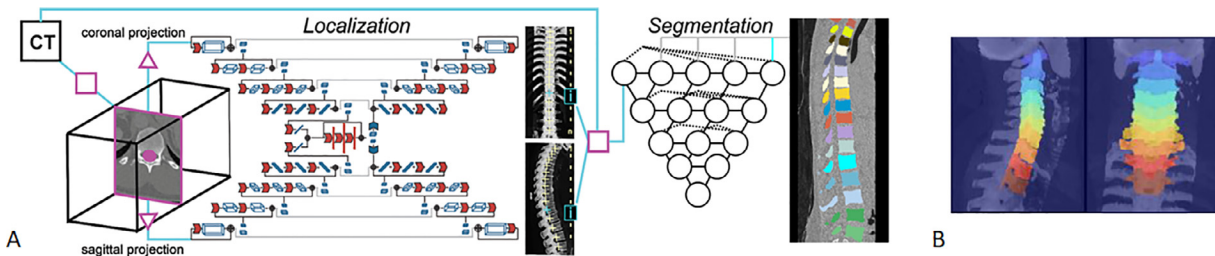


Fig. C.28. (A) An overview of the three-stage framework proposed by Netherton *T.*: Spinal canal segmentation, localisation, and segmentation. (B) Ground truth sagittal and coronal intensity projection image pairs used in the training of the second stage. Each colored planar projection is housed in a separate channel. Centroids of each colored mask provide coordinates used in subsequent stages in this approach.

resampled to have isotropic voxel sizes (1.0mm^3) and set to a common orientation. Ground truth localisation coordinates were not used in this approach. In total, 160 pairs of CT scans and segmentations were obtained and split into five groups for cross-validation.

Spinal canal segmentation. First, the spinal canal is segmented via a 2-dimensional FCN-8s with batch normalisation on the axial CT slices. Pairs of intensity projection images (sagittal and coronal) are then generated about a volume of interest (cropped from the CT scan) surrounding the spinal canal. These image pairs provide the network with sagittal and coronal views of the vertebral column and have a fixed width but variable length l (where l is the length of the CT scan); their corresponding ground-truth labels are then assigned individual channels (27 total) to account for each vertebral level. For the training stage, planar segmentation masks posterior to the spinal canal are removed to produce modified vertebral coordinates. In order to provide a large number of image augmentations, intensity projection image pairs (and corresponding ground truth masks) are incrementally cropped from the superior-inferior, inferior-superior, and medial-lateral directions.

Multi-view spinal localisation. X-Net (Netherton et al., 2020), the localisation architecture, inputs the sagittal and coronal intensity projection pairs and outputs labeled, multi-dimensional sagittal and coronal arrays of individual vertebral column segmentations. X-Net, inspired by Sekuboyina et al. (2020) and Milletari et al. (2016), incorporates residual connections, pReLU activations, and is end-to-end trainable. By combining centre-of-mass coordinates from sagittal and coronal planar segmentations, 3-dimensional locations are obtained for each vertebral body. During training, the loss function, which incorporated the soft-Dice loss and cross-entropy loss, was applied to each view (i.e. coronal and sagittal, $L = L_s + L_c$). Augmentations were applied during

training with a frequency of 0.7; coronal arrays were flipped left-right with a frequency of 0.5. Training was performed on a 16GB NVIDIA-V100 with batch size 8. Each model was trained for at least 26,000 iterations using early stopping.

Deeply supervised vertebral body segmentation. To perform vertebral body segmentation, a UNet++ architecture using skip connections, multi-class structure, and deep supervision is designed based on work by (Zhou et al., 2019). Using ground truth images and segmentations, three channel arrays are formed for each vertebral level by cropping around the centre of mass of each vertebral level. Separate channels contained background, adjacent vertebral levels, and the central vertebral level, respectively. For each 3-dimensional coordinate (from the second stage), the CT scan is cropped to form separate volumes of interest ($120 \times 96 \times 96\text{mm}^3$). The top two most supervised outputs from each prediction are averaged to yield the vertebral body of interest.

Paetzold J. et al.: A 2D-UNet on the VERSE data

The authors implement a 2-D segmentation architecture for slices of the sagittal orientation of the 3-D dataset using a 2D U-Net (Ronneberger et al., 2015). The encoder is made of a ResNet-34 backbone pre-trained on the ImageNet. The network is trained by optimising an equally weighted sum of the Dice loss and the binary cross-entropy loss (BCE) with data augmentations such as flipping, rotation, scaling, and shifting. The images are centre-cropped to 512 by 512 pixels to account for the irregular image sizes during training. All networks are implemented in Pytorch using the Adam optimiser and are trained for 1000 epochs. After prediction, the 2D slices are stacked together to reconstruct the 3D

volume. The training was carried out on an NVIDIA QUADRO RTX 8000 GPU with a batch size of 52.

Xiangshang Z. et al.: *Vertebra Labelling and Segmentation using the Btrfly-net and the nnU-net*

The authors design an improved Btrfly Network (Sekuboyina et al., 2018) to detect the key points of the vertebrae and then build an nn-Unet (Isensee et al., 2019) to segment the vertebral regions. Both the labelling and segmentation tasks are handled independently.

Vertebra Labelling. Similar to Sekuboyina et al. (2018), the authors work with 2D sagittal and coronal MIP. Improving on it, changes were made to the model architecture and the training procedure. Two convolution layers for each layer of encoder and decoder in the network followed by batch-normalisation and ReLU non-linearity after each convolution layer. Kaiming-initialisation is used for the network parameters. In terms of data-based enhancements, the authors use horizontal and vertical flip for augmentation and with normalisation.

Vertebra segmentation. The preprocessing and training procedure of the nnU-Net is retained. On top of it, data augmentation is applied on the fly during training using the batch-generators framework (Isensee et al., 2020). Specifically, elastic deformations, random scaling, and random rotations are used. If the data is anisotropic, the spatial transformations are applied in-plane as 2D transformations. Once trained, cases are predicted using a sliding window approach with half the patch size overlap between predictions.

Yeah T. et al.: *A Coarse-to-Fine Two-stage Framework for Vertebra Labeling and Segmentation.*

The author propose a two-stage network to achieve vertebra labeling and segmentation. Firstly, the low-resolution net determines the rough target location from downsampled CT images. Secondly, by feeding the first stage's prediction results (upsampling before feeding) and high-resolution CT scans into a full resolution net, more accurate vertebra classification and segmentation are achieved. Considering the competition among different vertebra classes especially for adjacent vertebra, finally connected component analysis is applied to refine vertebrae segmentation results.

The two-stage cascaded segmentation pipeline consists of two steps. Firstly a coarse location of spine RoI is obtained based on a lightweight low-resolution 3D U-Net from 3D CT scans with low resolution. Secondly the RoI and the accurate segmentation results are performed with a high-resolution 3D U-Net. Finally some post-processing methods are adopted to fill the holes inside each vertebrae and rule-based methods to recalibrate the vertebrae label. Both the low-resolution network and the high-resolution network have 26 output channels (C1-C7, T1-T13, L1-L6).

The first stage preprocesses the training 3D CT scans to a larger spacing through downsampling and train the low-resolution 3D U-Net model with a patch size of $224 \times 128 \times 96$. The second stage preprocesses 3D CT images to smaller spacing through upsampling and crops the RoI of spine regions as the training dataset for a high-resolution U-Net model with a patch size of $256 \times 96 \times 80$.

Preprocessing and Augmentation. All input images are normalised zero mean and unit standard deviation (based on foreground voxels only). The data augmentation include elastic deformation, rotation transformation, gamma transformation, random cropping, etc.

Loss and Optimisation. The low-resolution model with a classical combination of Dice loss and cross-entropy loss, while training the high-resolution model with a dynamic hybrid loss combining Dice loss and *weighted* cross-entropy loss. A model with a dynamic

hybrid loss combining Dice loss and Adam optimiser with an initial learning rate of 10^{-4} was used. During training, an exponential moving average of the validation and training losses is used. Whenever the training loss does not improve within the last 30 epochs, the learning rate is reduced by factor 5. The training is terminated automatically if validation loss does not improve within the last 50 epochs.

Zeng C.: *Two-stage Keypoint Location Pipeline for Vertebrae Location and Segmentation.*

The author proposes a two-stage keypoint detection pipeline for vertebral labeling based on the scheme of Payer et al. (2019) which uses Spatial-Configuration-Net and U-Net in VERSE'19 described in Section 3.2.

Additional Data and Preprocessing. An additional 13 data sets from the VERSE'19 training set are used. The data is first pre-processed to the RAI direction. Data augmentation includes rotation, intensity shift, scaling and elastic deformation. The model is trained with all 113 cases.

Localisation. To localize centres of the vertebrae, five keypoints location and global vertebrae location is performed separately. For the five keypoints, which contains the first and last two vertebral masses of the cervical spine, thoracic spine and lumbar spine, a network is designed of which the backbone is an HRNet (Sun et al., 2019) to regress the five keypoint heatmaps. The significance of the first stage is for better identification of several vertebral masses with obvious characteristics. The second stage follows Payer et al. (2019), with a re-designed channel attention block in the network with a weighted loss function.

Vertebrae Segmentation. For vertebrae segmentation, a binary segmentation network is trained based on the outcome of the labelling stage. A U-Net with an inputs size of $128 \times 128 \times 64$ is used. The loss function is a mixture of Dice loss and binary cross-entropy loss.

Zhang A. et al.: *A Segmentation-Based Framework for Vertebrae Localisation and Segmentation.*

In general, the vertebrae localisation and segmentation tasks are performed in a four-step approach: 1) spine localisation to obtain the region of interest, 2) single-class key point localisation to obtain the potential vertebrae candidates, 3) a triple-class vertebrae segmentation to obtain the individual mask and main category of each vertebrae, and 4) rule-based post-processing.

A variant of V-Net with a mixture of Dice and binary cross-entropy loss is utilised in the first three steps and only a few hyperparameters are changed in each step, such as input/output shape, depth, width, etc. Step 2 and step 3 could be corrected by each other in an 'intertwined' way as mentioned above: the proposed key-point candidates are used as input for step 3 to specify the vertebra to be segmented if the resulting segmentation result does not seem to be a mask (the volume is not large enough), then the proposed key-point can be regarded as a false positive.

Spine Localisation. To obtain the spinal centerline, a variant of V-Net is used to regress a heatmap of the spinal centerline. The input is a 4-time downsampled single-channel 3D-patch with a size of $64 \times 64 \times 64$. The sliding window approach is applied to serve the network with the specific size of local cubes. The heatmaps are generated using a Gaussian kernel by a kernel size (5, 5, 5) and sigma (6, 6, 6) on the downsampled mask to keep unique 3D connected domain. The output heatmap is converted to a binary mask by a threshold of 0.4 and resampled back to the origin image scale for later use.

Keypoint Localization. A similar variant of V-Net is employed to regress a heatmap of the spine. The input in this step is a single-

channel 3d-patch with the size of $64 \times 128 \times 128$. Sliding window approach is applied as above. The heatmaps are generated by kernel size (7, 9, 9) and sigma (6, 6, 6) on the original scale based on the JSON label to ensure they are independent and disconnected. The proposed regression results are converted to a binary mask by a threshold of 0.4 and the centroid is calculated for each cluster.

Vertebrae Segmentation. Considering half of the vertebrae account for a lack of samples for a 26-class classification, a triple-class segmentation task is defined to segment three categories: 'cspine', 'tspine' or 'lspine' for each vertebra in this step. A variant of V-Net is employed. The input is a cropped 3D patch around the localised centroid obtained from step 2.

Rule-based Post-processing. In this step a simple post-preprocessing logic is applied to create the final multi-label result. If more than one category of vertebrae is found in one case, the two or four 'split points' which is C7-T1 and T12-L1 can be localised. Then the others can be deduced based on these split points. If split points are not found in one case, then 'cspine' vertebrae are deduced from C1 to the bottom and 'lspine' ones are deduced from the bottom to the top.

References

- Angermann, C., Haltmeier, M., Steiger, R., Pereverzyev, S., Gizewski, E., 2019. Projection-based 2.5 d u-net architecture for fast volumetric segmentation. In: 2019 13th International conference on Sampling Theory and Applications (SampTA). IEEE, pp. 1–5.
- Anitha, D.P., Baum, T., Kirschke, J.S., Subburaj, K., 2020. Effect of the intervertebral disc on vertebral bone strength prediction: a finite-element study. *The Spine Journal* 20 (4), 665–671.
- Athertya, J.S., Kumar, G.S., 2016. Automatic segmentation of vertebral contours from ct images using fuzzy corners. *Comput. Biol. Med.* 72, 75–89.
- Bromiley, P.A., Kariki, E.P., Adams, J.E., Coates, T.F., 2016. Fully automatic localisation of vertebrae in ct images using random forest regression voting. In: International Workshop on Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 51–63.
- Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S., 2015. Multi-modality vertebrae recognition in arbitrary views using 3d deformable hierarchical model. *IEEE Trans Med Imaging* 34 (8), 1676–1693.
- Castro-Mateos, I., Pozo, J.M., Pereañez, M., Lekadir, K., Lazary, A., Frangi, A.F., 2015. Statistical interspace models (sims): application to robust 3d spine segmentation. *IEEE Trans Med Imaging* 34 (8), 1663–1675.
- Cauley, J., Thompson, D., Ensrud, K., Scott, J., Black, D., 2000. Risk of mortality following clinical fractures. *Osteoporosis International* 11 (7), 556–561.
- Chen, D., Bai, Y., Zhao, W., Ament, S., Gregoire, J., Gomes, C., 2020. Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In: International Conference on Machine Learning. PMLR, pp. 1500–1509.
- Chen, H., Shen, C., Qin, J., Ni, D., Shi, L., Cheng, J.C., Heng, P.-A., 2015. Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 515–522.
- Chen, J., Wang, Y., Guo, R., Yu, B., Chen, T., Wang, W., Feng, R., Chen, D.Z., Wu, J., 2019. Lsrc: A long-short range context-fusing framework for automatic 3d vertebra localization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 95–103.
- Chu, C., Belavý, D.L., Armbrrecht, G., Bansmann, M., Felsenberg, D., Zheng, G., 2015. Fully automatic localization and segmentation of 3d vertebral bodies from ct/mr images via a learning-based method. *PLoS ONE* 10 (11), e0143327.
- Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 424–432.
- Forsberg, D., Sjöblom, E., Sunshine, J.L., 2017. Detection and labeling of vertebrae in mr images using deep learning with clinical annotations as training data. *J Digit Imaging* 30 (4), 406–412.
- Frosio, I., Kautz, J., 2018. Statistical nearest neighbors for image denoising. *IEEE Trans. Image Process.* 28 (2), 723–738.
- Girshick, R., 2015. Fast r-cnn. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587.
- Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E., 2012. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*.
- Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A., 2013. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 262–270.
- Guan, Q., Wan, X., Lu, H., Ping, B., Li, D., Wang, L., Zhu, Y., Wang, Y., Xiang, J., 2019. Deep convolutional neural network inception-v3 model for differential diagnosing of lymph node in cytological images: a pilot study. *Ann Transl Med* 7 (14).
- Hammernik, K., Ebner, T., Stern, D., Urschler, M., Pock, T., 2015. Vertebrae Segmentation in 3D Ct Images Based on a Variational Framework. In: Recent advances in computational methods and clinical applications for spine imaging. Springer, pp. 227–233.
- Hanaoka, S., Nakano, Y., Nemoto, M., Nomura, Y., Takenaga, T., Miki, S., Yoshikawa, T., Hayashi, N., Masutani, Y., Shimizu, A., 2017. Automatic detection of vertebral number abnormalities in body ct images. *Int J Comput Assist Radiol Surg* 12 (5), 719–732.
- He, K., Gkioxari, G., Dollár, P., Girshick, R., 2018. Mask r-cnn. eprint 1703.06870.
- Howlett, D.C., Drinkwater, K.J., Mahmood, N., Illes, J., Griffin, J., Javid, K., 2020. Radiology reporting of osteoporotic vertebral fragility fractures on computed tomography studies: results of a uk national audit. *Eur Radiol.* <https://doi.org/10.1007/s00330-020-06845-2>.
- Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2019. Squeeze-and-excitation networks. eprint 1709.01507.
- Ibragimov, B., Korez, R., Likar, B., Pernuš, F., Vrtovec, T., 2015. Interpolation-based Detection of Lumbar Vertebrae in Ct Spine Images. In: Recent Advances in Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 73–84.
- Ibragimov, B., Korez, R., Likar, B., Pernuš, F., Xing, L., Vrtovec, T., 2017. Segmentation of pathological structures by landmark-assisted deformable models. *IEEE Trans Med Imaging* 36 (7), 1457–1469.
- Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T., 2014. Shape representation for efficient landmark-based segmentation in 3-d. *IEEE Trans Med Imaging* 33 (4), 861–874.
- Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., et al., 2020. batchgenerators-a python framework for data augmentation. 2020.
- Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2019. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128.
- Jakubicek, R., Vicar, T., Chmelik, J., 2020. A tool for automatic estimation of patient position in spinal ct data. In: European Medical and Biological Engineering Conference. Springer, pp. 51–56.
- Janssens, R., Zeng, G., Zheng, G., 2018. Fully automatic segmentation of lumbar vertebrae from ct images using cascaded 3d fully convolutional networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 893–897.
- Kadoury, S., Labelle, H., Paragios, N., 2011. Automatic inference of articulated spine models in ct images using high-order markov random fields. *Med Image Anal* 15 (4), 426–437.
- Kadoury, S., Labelle, H., Paragios, N., 2013. Spine segmentation in medical images using manifold embeddings and higher-order mrf. *IEEE Trans Med Imaging* 32 (7), 1227–1238.
- Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging* 29 (1), 196–205.
- Klinder, T., Ostermann, J., Ehm, M., Franz, A., Kneser, R., Lorenz, C., 2009. Automated model-based vertebra detection, identification, and segmentation in ct images. *Med Image Anal* 13 (3), 471–482.
- Korez, R., Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T., 2015. A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation. *IEEE Trans Med Imaging* 34 (8), 1649–1662.
- Korez, R., Likar, B., Pernuš, F., Vrtovec, T., 2016. Model-based segmentation of vertebral bodies from mr images with 3d cnns. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 433–441.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 25, 1097–1105.
- Laouissat, F., Sebaaly, A., Gehrchen, M., Roussouly, P., 2018. Classification of normal sagittal spine alignment: refounding the roussouly classification. *European Spine Journal* 27 (8), 2002–2011.
- Lessmann, N., van Ginneken, B., Išgum, I., 2018. Iterative convolutional neural networks for automatic vertebra identification and segmentation in ct images. In: *Medical Imaging 2018: Image Processing*, Vol. 10574. International Society for Optics and Photonics, p. 1057408.
- Lessmann, N., van Ginneken, B., de Jong, P.A., Išgum, I., 2019. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Med Image Anal* 53, 142–155. doi:10.1016/j.media.2019.02.005.
- Leventon, M.E., Grimson, W.E.L., Faugeras, O., 2002. Statistical shape influence in geodesic active contours. In: 5th IEEE EMBS International Summer School on Biomedical Imaging, 2002.. IEEE, pp. 8–pp.
- Li, Y., Cheng, X., Lu, J., 2018. Butterfly-net: optimal function representation based on convolutional neural networks. arXiv preprint arXiv:1805.07451.
- Liao, H., Mesfin, A., Luo, J., 2018. Joint vertebrae identification and localization in spinal ct images by combining short-and long-range contextual information. *IEEE Trans Med Imaging* 37 (5), 1266–1275.
- Liebl, H., Schinz, D., Sekuboyina, A., Malagutti, L., Löffler, M.T., Bayat, A., Husseini, M.E., Tetteh, G., Grau, K., Niederreiter, E., et al., 2021. A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data. arXiv preprint arXiv:2103.06360.
- Lim, P.H., Bağcı, U., Bai, L., 2014. A Robust Segmentation Framework for Spine Trauma Diagnosis. In: *Computational Methods and Clinical Applications for Spine Imaging*. Springer, pp. 25–33.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal loss for dense object detection. eprint 1708.02002.

- Löffler, M., Sollmann, N., Mei, K., Valentinitzsch, A., Noël, P., Kirschke, J., Baum, T., 2020. X-Ray-based quantitative osteoporosis imaging at the spine. *Osteoporosis International* 1–18.
- Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A.-L., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S., 2020. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence* 2 (4), e190138. doi:10.1148/ryai.2020190138.
- Mader, A.O., Lorenz, C., von Berg, J., Meyer, C., 2019. Automatically localizing a large set of spatially correlated key points: A case study in spine imaging. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 384–392.
- Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun* 9 (1), 1–13.
- Major, D., Hladůvka, J., Schulze, F., Bühler, K., 2013. Automated landmarking and labeling of fully and partially scanned spinal columns in ct images. *Med Image Anal* 17 (8), 1151–1163.
- Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans Med Imaging* 34 (10), 1993–2024.
- Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE, pp. 565–571.
- Müller, D., Bauer, J.S., Zeile, M., Rummeny, E.J., Link, T.M., 2008. Significance of sagittal reformations in routine thoracic and abdominal multislice ct studies for detecting osteoporotic fractures and other spine abnormalities. *Eur Radiol* 18 (8), 1696–1702.
- Netherton, T.J., Rhee, D.J., Cardenas, C.E., Chung, C., Klopp, A.H., Peterson, C.B., Howell, R.M., Balter, P.A., Court, L.E., 2020. Evaluation of a multiview architecture for automatic vertebral labeling of palliative radiotherapy simulation ct images. *Med Phys* 47 (11), 5592.
- Oxland, T.R., 2016. Fundamental biomechanics of the spine—what we have learned in the past 25 years and future directions. *J Biomech* 49 (6), 817–832.
- Payer, C., Štern, D., Bischof, H., Urschler, M., 2019. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Med Image Anal* 54, 207–219.
- Payer, C., Štern, D., Bischof, H., Urschler, M., 2020. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. In: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, Vol. 5, pp. 124–133. doi:10.5220/0008975201240133.
- Pereañez, M., Lekadir, K., Castro-Mateos, I., Pozo, J.M., Lazáry, Á., Frangi, A.F., 2015. Accurate segmentation of vertebral bodies and processes using statistical shape decomposition and conditional models. *IEEE Trans Med Imaging* 34 (8), 1627–1639.
- Rasoulia, A., Rohling, R., Abolmaesumi, P., 2013. Lumbar spine segmentation using a statistical multi-vertebrae anatomical shape+ pose model. *IEEE Trans Med Imaging* 32 (10), 1890–1900.
- Redmon, J., Farhadi, A., 2016. Yolo9000: Better, faster, stronger. eprint 1612.08242.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*.
- Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115 (3), 211–252.
- Seim, H., Kainmueller, D., Heller, M., Lamecker, H., Zachow, S., Hege, H.-C., 2008. Automatic segmentation of the pelvic bones from ct data based on a statistical shape model. pp. 93–100. doi:10.2312/VCBM/VCBM08/093-100.
- Sekuboyina, A., Kukačka, J., Kirschke, J.S., Menze, B.H., Valentinitzsch, A., 2017. Attention-driven deep learning for pathological spine segmentation. In: *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer, pp. 108–119.
- Sekuboyina, A., Rempfler, M., Kukačka, J., Tetteh, G., Valentinitzsch, A., Kirschke, J.S., Menze, B.H., 2018. Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018*.
- Sekuboyina, A., Rempfler, M., Valentinitzsch, A., Menze, B.H., Kirschke, J.S., 2020. 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.
- Sekuboyina, A., Valentinitzsch, A., Kirschke, J.S., Menze, B.H., 2017. A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets. arXiv preprint arXiv:1703.04347.
- Shamonin, D.P., Bron, E.E., Lelieveldt, B.P., Smits, M., Klein, S., Staring, M., 2014. Fast parallel image registration on cpu and gpu for diagnostic classification of alzheimer's disease. *Front Neuroinform* 7, 50.
- Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Štern, D., Likar, B., Pernuš, F., Vrtovec, T., 2011. Parametric modelling and segmentation of vertebral bodies in 3d ct and mr spine images. *Physics in Medicine & Biology* 56 (23), 7505.
- Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703.
- Suzani, A., Rasoulia, A., Seitel, A., Fels, S., Rohling, R.N., Abolmaesumi, P., 2015. Deep learning for automatic localization, identification, and segmentation of vertebral bodies in volumetric mr images. In: *Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*, Vol. 9415. International Society for Optics and Photonics, p. 941514.
- Suzani, A., Seitel, A., Liu, Y., Fels, S., Rohling, R.N., Abolmaesumi, P., 2015. 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*. Springer, pp. 678–686.
- Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 15 (1), 1–28.
- Wigh, R.E., 1980. The thoracolumbar and lumbosacral transitional junctions. *Spine* 5 (3), 215–222.
- Williams, A.L., Al-Busaidi, A., Sparrow, P.J., Adams, J.E., Whitehouse, R.W., 2009. Under-reporting of osteoporotic vertebral fractures on computed tomography. *Eur J Radiol* 69 (1), 179–183.
- Wu, Y., He, K., 2018. Group normalization. eprint 1803.08494.
- Yang, D., Xiong, T., Xu, D., Huang, Q., Liu, D., Zhou, S.K., Xu, Z., Park, J., Chen, M., Tran, T.D., et al., 2017. 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*. Springer, pp. 633–644.
- Yang, D., Xiong, T., Xu, D., Zhou, S.K., Xu, Z., Chen, M., Park, J., Grbic, S., Tran, T.D., Chin, S.P., et al., 2017. 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*. Springer, pp. 498–506.
- Yao, J., Burns, J.E., Forsberg, D., Seitel, A., Rasoulia, A., Abolmaesumi, P., Hamernik, K., Urschler, M., Ibragimov, B., Korez, R., Vrtovec, T., Castro-Mateos, I., Pozo, J.M., Frangi, A.F., Summers, R.M., Li, S., 2016. A multi-center milestone study of clinical vertebral ct segmentation. *Computerized Medical Imaging and Graphics* 49, 16–28. doi:10.1016/j.compmedimag.2015.12.006.
- Yao, J., Burns, J.E., Munoz, H., Summers, R.M., 2012. Detection of vertebral body fractures based on cortical shell unwrapping. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 509–516.
- Yu, Q., Yang, D., Roth, H., Bai, Y., Zhang, Y., Yuille, A.L., Xu, D., 2020. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4126–4135.
- Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-iou loss: Faster and better learning for bounding box regression. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 12993–13000.
- Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans Med Imaging* 39 (6), 1856–1867.



Anduin: An open-source, web-based spine segmentation tool



Anduin: An open-source, web-based spine segmentation tool

Largely based on the work presented in this work, along with the supporting works on vertebrae segmentation, we developed *anduin*, an publicly-available, web-based spine segmentation tool. *anduin* is hosted at Klinikum rechts der Isar and was developed by Giles Tetteh, Malek Husseini, and **Anjany Sekuboyina**, under the supervision of Dr. Jan S. Kirschke. It encompasses our experience in spine image processing through multiple publications as well as from the largest spine segmentation benchmark yet [4].

Overview

The landing page of *anduin*, hosted at anduin.bonescreen.de is shown in Fig. F.1. As input, *anduin* takes a spine CT image in the NIFTI file format along with an optional JSON side-car consisting of the DICOM header information (cf. Fig. F.2). In this image, the following steps are performed: spine localisation, vertebrae labelling, vertebrae segmentation, and vertebral subregion¹ segmentation. The processing of a scan in progress is shown in Fig. F.3. The output of each of these stages can be downloaded as shown in Fig. F.4. Finally, an example out of the tool is shown in Fig. F.5, indicating the vertebral centroids and segmentation masks of both vertebrae and its subregions.

Contributions

Project inception and planning, prototyping, development of machine-learning backend, project supervision (technology) and management.

¹Subregions: every vertebrae is further segmented into ten subregions such as vertebral arch, spinous process, cortex, left and right transverse process etc.

F. ANDUIN: AN OPEN-SOURCE, WEB-BASED SPINE SEGMENTATION TOOL

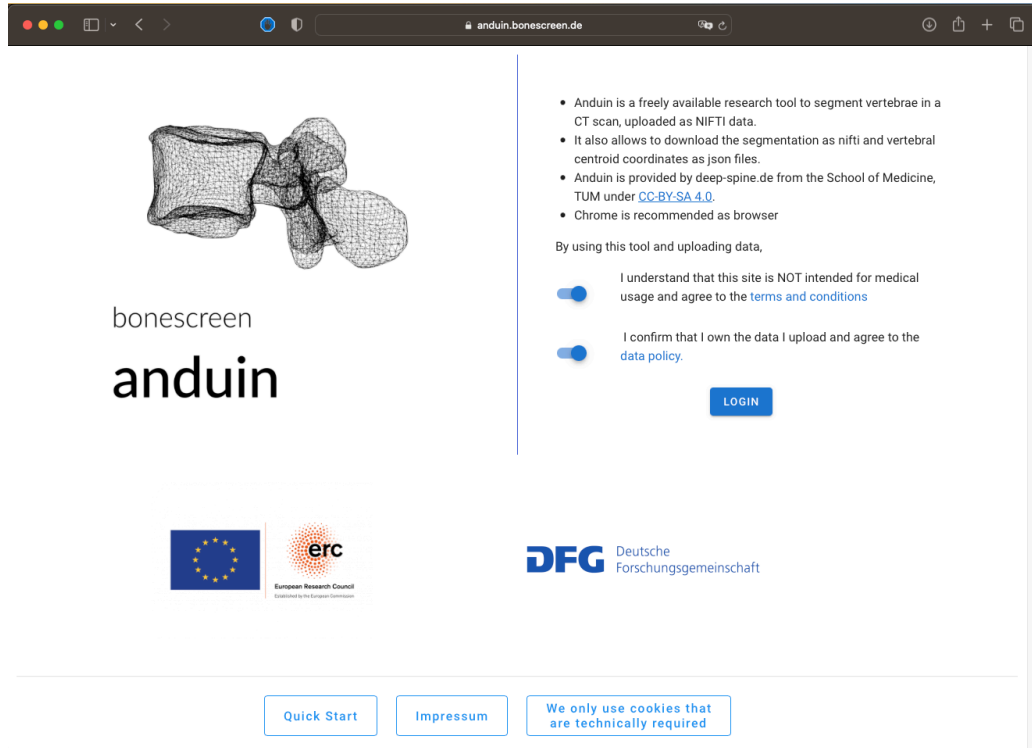


Figure F.1: Anduin’s landing page: Anduin is hosted at anduin.bonescreen.de, released under CC-BY-SA 4.0 license. Users have the ability to request for user accounts after agreeing to the data policy along with anduin’s terms and conditions on data processing

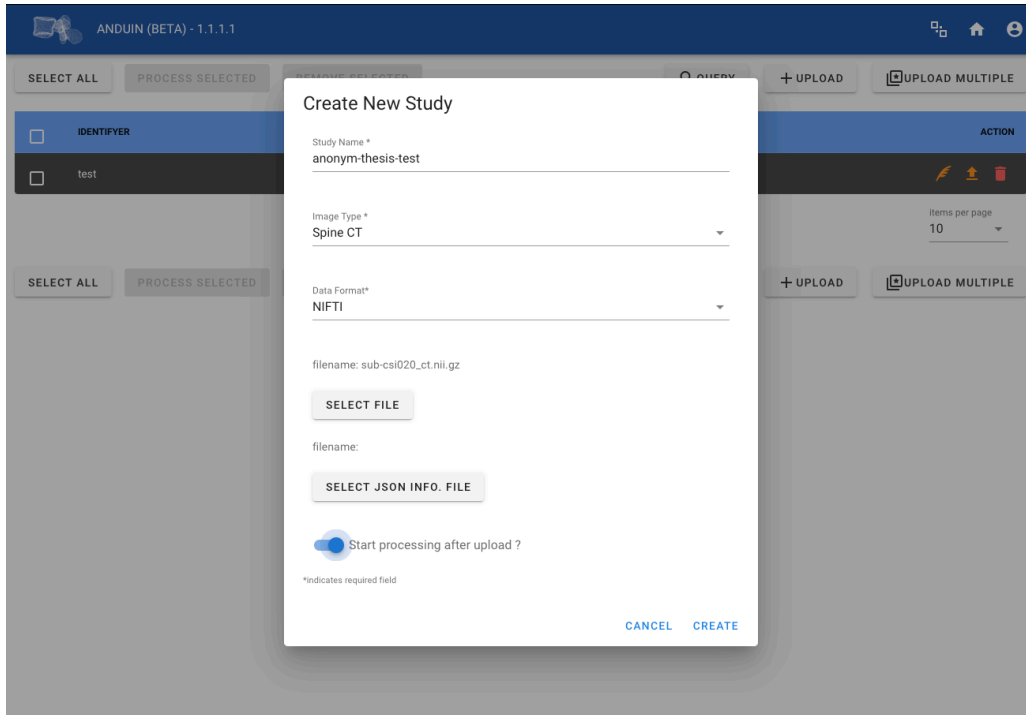


Figure F.2: Upload dialogue: Users have the ability to upload NIFTI files (*.nii or *.nii.gz) along with a JSON side car containing the header information




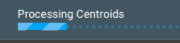






<input type="checkbox"/>	IDENTIFIER	CREATED	LAST CHANGES	STATUS	ACTION
<input checked="" type="checkbox"/>	anonym-thesis-test	24/04/2023, 16:50:05	24/04/2023, 16:50:05	CT ax 2mm	  
<input type="checkbox"/>	BMD Evaluation	24/04/2023, 16:50:05	24/04/2023, 16:50:05	Processing Centroids 	  
<input type="checkbox"/>	test	21/09/2022, 15:34:01	21/09/2022, 15:34:28	CT sag 2mm	  

Figure F.3: Processing status: Once uploaded, the scan can be processed to result in an ‘evaluation’ containing the processing artefacts. Processing of a CT of diagnostic quality takes approximately 1–2 minutes.

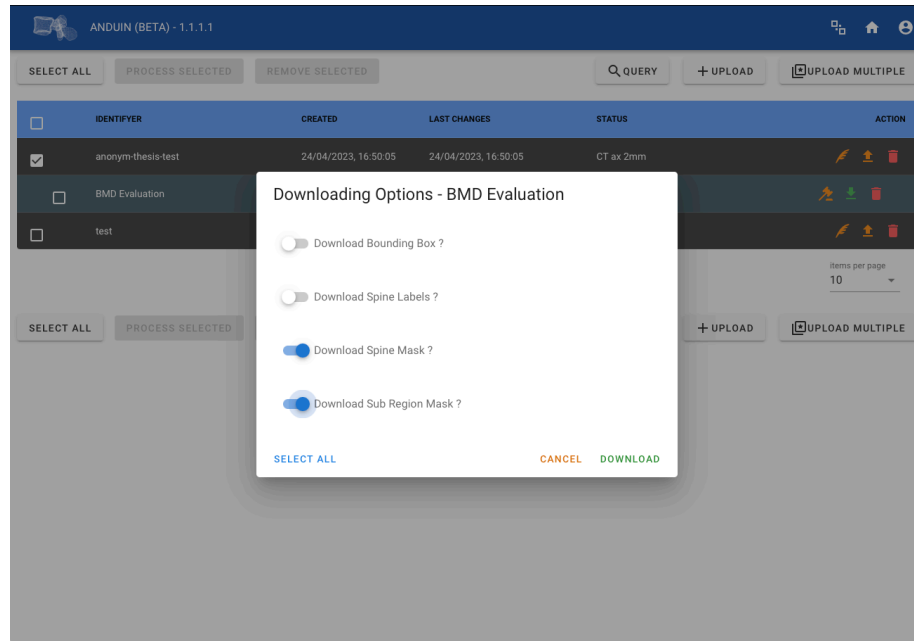


Figure F.4: Artefact download: Once processed, every artefact of the processing pipeline can be download, as required. This includes the bounding box around the spine, the vertebral centroids, the segmentation masks or the vertebrae and their subregion masks.

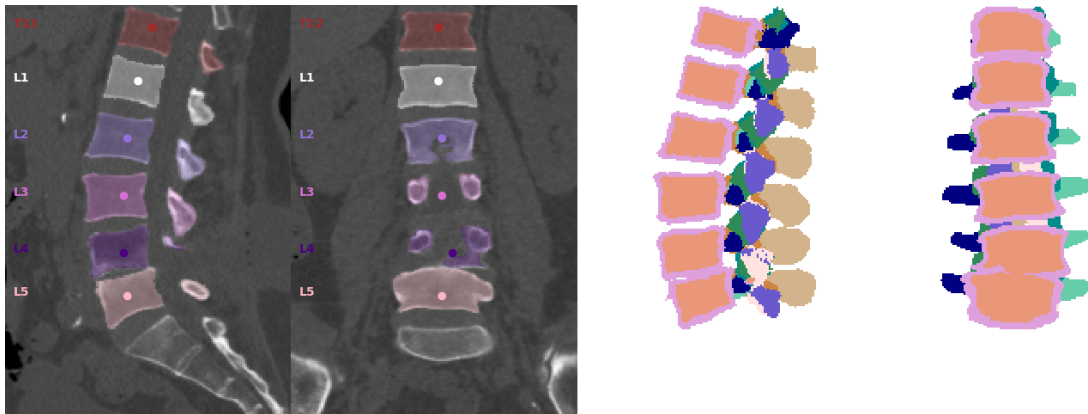


Figure F.5: A snapshot of *anduin*'s output, the left two tiles showing the sagittal and coronal reformations with the predicted centroids and segmentation masks and the two tiles on the right showing the maximum intensity projection of the subregion mask (sagittal and coronal).