Attention-driven Deep Learning for Pathological Spine Segmentation

Anjany Sekuboyina[1,2,⋆], Jan Kukačka[1,2,⋆], Jan S. Kirschke[2], Bjoern H. Menze[1], and Alexander Valentinitsch[1,2]

[1]Department of Informatics, Technische Universität München, Munich, Germany
[2]Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, Munich, Germany
anjany.sekuboyina@tum.de

Abstract. Accurate segmentation of the spine in CT images is mandatory for quantitative analysis, e.g. in osteoporosis, but remains challenging due to high variability in vertebral morphology and spinal anatomy among patients. Conventionally, spine segmentation was performed by model-based techniques employing spine atlases or statistical shape models. We argue that such approaches, even though intuitive, fail to address clinical abnormalities such as vertebral fractures, scoliosis, etc. We propose a novel deep learning-based method for segmenting the spine, which does not rely on any pre-defined shape model. We employ two networks: one for localisation and another for segmentation. Since a typical spine CT scan cannot be processed at once owing to its large dimensions, we find that both nets are essential to work towards a perfect segmentation. We evaluate our framework on three datasets containing healthy and fractured cases: two private and one public. Our approach achieves a mean Dice coefficient of ∼0.87, which is comparable but not higher than the state-of-art model-based approaches. However, we show that our approach handles degenerate cases more accurately.

1 Introduction

Spine segmentation is a crucial component in quantitative medical image analysis. It directly allows detection and assessment of vertebral fractures and indirectly supports modelling and monitoring of the spinal ageing process. In this work we propose a method based on 'deep-learning', that generates precise spine segmentations on computed tomography (CT) images. It overcomes the drawbacks of earlier segmentation approaches and thus can be used in clinical settings. Particularly, our approach is capable of handling scans with varying fields-of-view (FOV) and degenerate spine or vertebrae.

Previous works often deal with spine segmentation in a multi-step approach incorporating spine localisation and vertebra detection followed by the segmentation [1, 2]. Various traditional computer vision techniques have been successfully

[⋆] Both the authors have contributed equally.

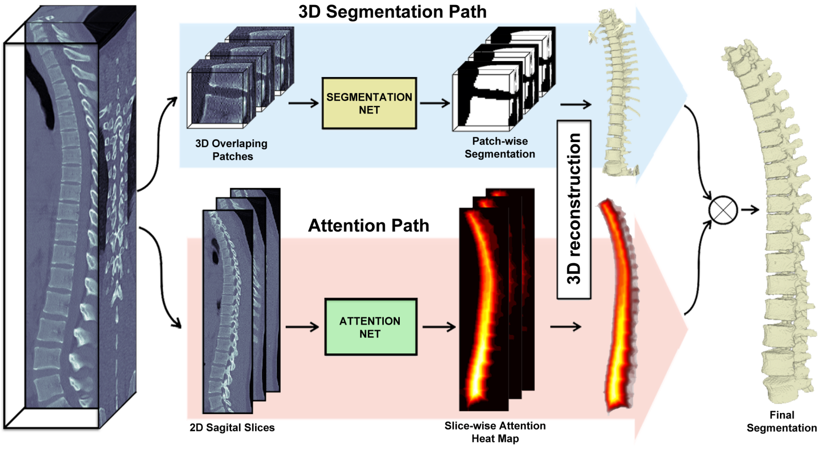

Fig. 1: Schematic outline of the proposed approach.

applied, such as active shape models and snake-based methods [3, 4], level sets[5] or graph-based approaches using normalized-cuts [6]. Most of these methods rely on prior knowledge in the form of spine atlases or statistical shape models which are used to provide a good initialisation. Such models reach state-of-the-art performance on healthy spines with no signs of osteoporotic fractures, attaining Dice coefficients (DICE) of over 0.9. However, osteoporotic patients often suffer from severe vertebral fractures in various stages and spinal deformities such as scoliosis. In such cases, model-dependent segmentation might fail due to the high variability of the unique shape of a fracture or deformity that does not resemble a mean shape model. A shape model is also restricted by its mesh interpolation algorithm, which makes extreme deformations unfeasible. Moreover, CT images acquired for preoperative planning due to other diseases in the thoracic or abdominal area have the spine in them as a consequence. Such opportunistic scans have varying FOVs, spatial resolution, and image reconstruction, in addition to variations in scan enhancements due to contrast agents. Model-based approaches, which rely on good initialisations, could fail in such cases either due to lack of landmarks for registration, uneven intensities, and noise. This calls for data-driven approach based on supervised learning that does not rely on pre-defined models, but learns the variability by training on several kinds of contingencies.

Machine learning-based approaches have proven to fulfil these requirements, given that enough data is available for their training. Glocker et *al.* [7] and Suzani et *al.* [8] attempt the vertebra detection problem on arbitrary FOVs using random forests and multi-layer perceptrons respectively. More recently, Chen et *al.* [9] try to make use of the omni-present convolutional neural networks (CNN), with a clever cost formulation, to detect vertebrae. Eventually the CNNs have gained large popularity also for image segmentation through the concept of fully-convolutional networks (FCN) allowing pixel-to-pixel training and inference on (nearly) arbitrary sized inputs [10]. The standard FCN architecture of a contracting and an expanding path with shortcut connections is exploited in recent works on segmentation in the context of medical imaging [11–13]. However, these approaches cannot be directly extended to obtain a dense segmentation of a spine scan due to the sheer spatial resolution of a scan. For instance, the segmentation net used in [10] works on inputs containing $\sim 2.5 \times 10^5$ pixels; a typical whole spine CT scan is about 100–1000 times larger, thereby making a straightforward extension of an FCN non-viable.

We combine the FCN architecture with a domain-specific data-preprocessing pipeline and data-augmentation scheme to propose a robust and scalable framework for spine segmentation in CT images. The method builds on the following key elements:

1. A low-resolution attention FCN for spine localisation that works on two-dimensional sagittal slices of a scan.
2. A high-resolution segmentation FCN for fine segmentation that takes three-dimensional patches as input.
3. A smart patch extraction strategy to incorporate the FOV invariance and bypass memory limitations.
4. A domain-specific data augmentation to increase the training set size and incorporate the typical biological variance.

Our approach is free from predefined shape models and is purely data-driven, and is thus highly generalisable across varying FOVs, spinal deformities, and spatial resolutions given sufficiently diverse training data. Methodological details are presented in Section 2. We evaluate our method on a large private dataset of 56 (a) healthy and (b) fractured patients. We also compare our approach against the state-of-the-art methods on a publicly available dataset from the 2014 MICCAI workshop on Computational Spine Imaging (CSI) [14]. Our approach achieves a comparable mean DICE of around 87%, while perfectly segmenting fine details of normal as well as deformed vertebrae. Details of the experiments follow in Section 3.

2 Methodology

We present our approach to spine segmentation in two stages. Firstly, a 2D FCN which provides a low-resolution localisation of the spine. Secondly, we present the 2D-3D FCN that generates high-resolution binary segmentations. Fusing the predictions of both the networks results in a good segmentation of a spine volume. An overview of our approach is shown in Figure 1.

2.1 Localization: Attention-Net

We exploit the the structure and position of the spine in a scan, that are generally invariant, to obtain a rough localisation of the spine. The network performing

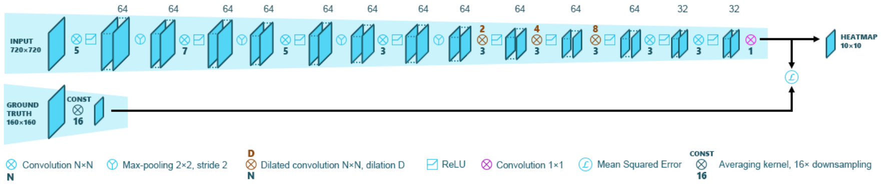

Fig. 2: **Attention-net**. The network utilizes 10 convolutional layers with stride 1, kernel sizes and dilation factors are denoted in the image. Moreover, the first 4 conv-layers are followed by max-pooling. Numbers above blobs represent number of features in the hidden feature space. Observe the drop in spatial resolution from the input to the output slice, thereby providing a low-resolution attention map for a given sagittal slice. Every nth sagittal slice from the scan volume is considered for inference, thereby reducing the number of forward passes for scan, which makes the attention generation very fast. Consequently, the original scan resolution is restored by interpolation in all three directions.

this task is called the *attention net*. Since sagittal view provides significant context on the spine's location, the attention net operates fully in 2D on sagittal slices. The net is fully convolutional and outputs a 2D map of lower dimension than the input. Every value ($\in [0, 1]$) in the predicted 2D map corresponds to a 16×16 region in the input, and represents the percentage of foreground voxels ('spine' voxels) in that region. Figure 2 illustrates the architecture of the *attention-net*. A 2D patch of 160×160 (padded to size 720×720) predicts a maps of size 10×10. Since context is of utmost importance for determining the presence of spine, we increase the receptive field with dilated convolutions [15] in the downstream convolutional layers. We incorporate the resolution reduction in the third dimension by working only on every n^{th} sagittal slice. At the end, given an input volume, the attention net works on sagittal slices and predicts a lower-resolution volume (called the *attention map*) whose values indicate the presence of spine. The attention map is then up-sampled to the input dimension for further use.

Training and Inference The ground truth for training is obtained from the available spine segmentations. Every volume is down-sampled by a factor of $16 \times 16 \times n$ ($n = 8$, in our experiments), each voxel representing the ratio of spine-voxels to total-voxels in its corresponding 16×16 region. The network is trained to minimise the mean-squared error between predicted map and ground truth. During training, given a scan volume, we train on large patches randomly sampled from sagittal slices with data augmentation through rigid transformations (2D rotations by ±20° and scaling of the axes by ±40%); we advocate the use of patches as incorporation of invariance to arbitrary FOVs. During test time, the patches are sampled from every n^{th} sagittal slice with overlap such that the entire slice is covered. The predicted low-resolution attention map for each of

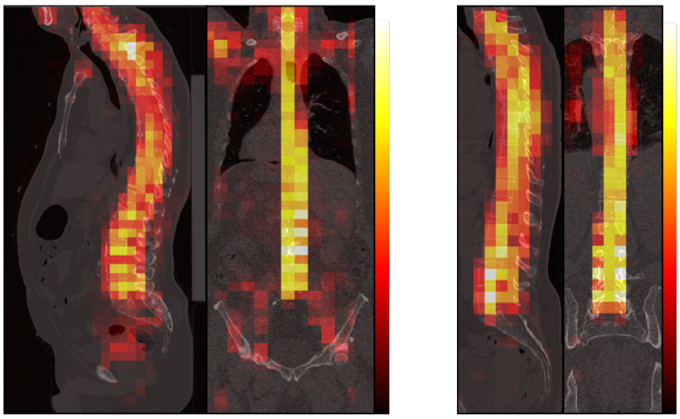

Fig. 3: **Attention maps**. (a) and (b) show two CT scans overlaid with the response of the Attention-Net aggregated over all slices in the direction of view. Notice that the response is focused predominantly on the spine. This attention-map is Gaussian smoothened, thresholded, and converted to a binary mask, which is then fused with the Segmentation-Net's response.

these patches is up-sampled to the resolution of the scan volume, and filtered with a 3D Gaussian kernel, $\mathcal{N}(0, \Sigma)$. Figure 3 shows an attention map of two test cases. For better visualisation of the 3D map, the response is aggregated in the direction of view and overlaid on the mid-slice. Observe that the net succeeds in localising the spine. This attention-map is thresholded and converted to a binary mask, which is then fused with the Segmentation-Net's response as elaborated in the following sections. The threshold-value and the covariance of the Gaussian smoothing (Σ) are tuned on the validation set.

2.2 Spine Segmentation: Segmentation-Net

Precise segmentation can be obtained when the receptive field of the network if small enough that it focuses on minute details, while being large enough to capture sufficient context. We achieve this by incorporating a patch-based approach. Such an approach also alleviates the restriction that the limited memory of a GPU imposes on the volume that can be processed by the network. We propose a segmentation net that is fully convolutional and a combination of 2D and 3D convolutions, building on the versatile 'U' architecture commonly used for segmentation [16, 12]. A detailed view of the network's architecture is shown in Figure 4. The input to the network is a 3D block from the scan, having larger receptive field in the sagittal view (for example, an input block could be of size 188×188×12, with the first two dimensions corresponding to the sagittal view). At an isotropic resolution of 1mm^3, the receptive field for predicting one voxel's

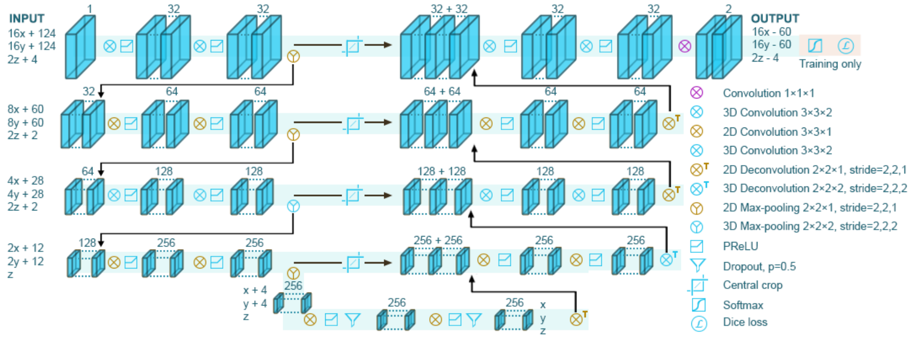

Fig. 4: **Segmentation-net**. The network has a five-level downsampling path, each level consisting of two convolutions with a ReLU activation and a max-pooling, and a symmetric upsampling path, which are connected by skip connections for recovery of high resolution. Notice the utilisation of a combination of 2D and 3D convolutions to process higher information in the sagittal direction, while conglomerating information from the adjacent slices. The size of the inputs and outputs is parameterised by the dimensions of the smallest blob on the path as x, y $\geq$ 4, z $\geq$ 3

label has a size of $\sim 18.4 \times 18.4 \times 0.8$ cm^3. The output is a dense pixel-wise segmentation of dimensions equal to that of the input.

Training and Inference We use a segmentation-centric loss function of DICE [13] as the objective function. At the training time we randomly sample patches from the input volumes and apply online generated rigid deformations to them as a form of data augmentation. Specifically, we apply rotations along the sagittal plane by $\pm 2°$ and scaling by $\pm 10\%$ along with minute contrast adjustments. Employing dropout after the convolutions on the lowest level of the contracting path gave a significant performance improvement. At the test time, the patches are extracted uniformly with overlap such that the resulting segmentations cover the entire test scan.

To conclude the approach: given a scan volume, the segmentation-net provides its dense segmentation. Since the latter splits the volume in sub-blocks, we observe that the context in these blocks is insufficient for perfect segmentation. We therefore observe several false positives in the form of *stray* segmentation, wherein, in addition to spine, other regions such as parts of the rib cage, pelvic bone, sacrum etc. are also segmented. The thresholded binary attention-map, which localises only the spine, is used as a mask over the segmentation-net's output to clear away the stray segmentations[1], resulting in the final segmentation map.

[1] A cascaded fusion of these nets was also tried where the patches for the segmentation net are obtained only from the region proposed by the attention map. We observed that this approach's accuracy was not superior to our approach of late-fusion.

3 Experiments and Results

Implementation details The two networks in our approach are implemented as standalone modules. Given a scan, the inference in both the nets run in parallel, and their responses are combined. As a preprocessing step, the CT volumes are subjected to anisotropic diffusion filtering in order to smoothen the homogeneous regions and improve the details around the edges. Both the nets were implemented in Caffe [17]. Adam solver was employed for optimising the loss. The nets were implemented on an Nvidia Titan X GPU with 12GB vRAM and trained till convergence with initial learning rates of 10^{-5} for segmentation net and 10^{-4} for the attention net. The convergence was faster for the attention net owing to its significantly lower number of parameters.

Data We evaluate our framework on three datasets: (Dataset 1) forty five patients without fractures but varying age from 25 to 69 years, (Dataset 2) eleven cases with vertebral fractures, to evaluate the performance on deformed vertebrae, and (Dataset 3) the spine segmentation challenge of the 2014 MICCAI workshop on Computational Spine Imaging dataset containing twenty CT scans, for comparison with other model-based techniques.

Datasets 1 and 2 are private in-house datasets gathered from our PACS. The scans were acquired over a period of two years for various patient examinations and not specifically for the spine analysis, which results in a high diversity in terms of patient-age, abnormalities, FOV, and scanner calibrations. The ground truth segmentation for this data was obtained by first using the approach in [1][2] and then manually corrected by a medical expert. Experiments on these datasets are therefore close to the clinical scenario.

Dataset 1 This is employed to validate the generalisability towards varying BMDs and scanner calibrations. It includes *healthy controls* (HC) with no fractures, which have been acquired with varying scanner settings (fields-of-view, spatial resolution, etc.). Thus, there is a significant variation in the Hounsfield units (HU) of the scans. We reserve three volumes as a validation set and another three volumes for testing, the remaining thirty nine are used for training.

Dataset 2 We utilise this to evaluate the performance on fracture cases (Fx). For this experiment we fused the forty five cases of Dataset 1 with eleven cases from this dataset, all of which have fractures. On top of the splits defined for Dataset 1, we add six fractured cases to the training set, two to the validation set, and two to the test set.

Dataset 3 As a final experiment, we intend to compare our algorithm with the other best-performing segmentation algorithms. For this task, we choose the public dataset of the segmentation challenge in CSI Workshop 2014. This contains

[2] We would like to thank Klinder et al., the authors and our industry partners (Philips, Hamburg), for providing us with the segmentation of Datasets 1 and 2 based on their approach in [1]

a total of twenty scans, ten for training phase and ten for testing. Of the test set, five scans are of healthy subjects less than 35 years old, and five other scans are of osteoporotic spines aged above 55 years. From the ten training volumes we reserve one for validation and use the remaining nine for fine-tuning of the model pre-trained on the Dataset 2. We refer the reader to [14] for a detailed description of the dataset.

Results The results of the experiments on the Datasets 1 and 2 are reported in the Table 1. Our method reaches good results both for healthy and fractured cases, proving to generalise well to a wide range of data. In both cases we observed the segmentation net suffer from stray segmentations which were mostly filtered out by the attention net, steadily improving the DICE by 1–10%. The resulting segmentation of our method are visualised in Figure 5, comparing it to a well-known model-based segmentation approach [1]. Notice the over-segmentation of the vertebral process regions (top row, HC) in the model-based approach. This is expected as the process of a vertebra has very high variability from patient-to-patient. Such a variability cannot be captured by an atlas or a shape model. Such over-segmentation does not occur in our approach. The bottom row illustrates our approaches performance on a fractured vertebra. Observe how the model-based segmentation fails to capture the deformities. This illustrates the restriction the interpolation algorithm in such approaches fails to capture extreme deviations from the mean shape. Our approach, however, successfully segments the degenerate vertebra as it learns to segment on the edges and is not hindered by any shape priors. We attribute the bleeding in the vertebral process regions (Figure 5, bottom row) to the predicted mask of the neighbouring slices.

The results of the experiment on Dataset 3 are reported in Table 2. We compare our approach with two methods that have been deemed best performing according to the challenge organisers. [4] uses a mean shape model based strategy followed by an efficient interpolation theory oriented mesh deformation, and performed best on healthy cases. [2] performed well on osteoporotic cases. It uses a multi-atlas based segmentation followed by a B-spline relaxation to adapt for the variability of vertebral structures. Our algorithm achieves a performance comparable to [2] on the healthy cohort. However, we observe that the performance on the osteoporotic test cases is not up to the mark; we identify the cause to be following: The fractured vertebra in the test set are treated by cement injection inside the vertebra (for example, case 26 and case 30 have cementing in T12 and L3 vertebrae). This procedure causes distinctive artefacts in the image, as opposed to untreated fractures. As there was not a single case of a cement-treated fracture in our training set and since our approach is purely data-driven, our method doesn't perform as expected. This can be observed by observing a more detailed view on the osteoporotic set's performance metric, where we obtain a highest Dice score of 89.4% and a median score of 85.48%. However, this performance can be easily improved by adding more representative data into the training set.

Dataset	DICE
Dataset 1 (HC)	87.6 ± 5.0
Dataset 2 (HC+Fx)	85.91 ± 4.8

Table 1: DICE (in %) for Datasets 1 and 2. The performance of our method is consistent among HC and Fx cases.

Approach	DICE	
	HC	Osteo.
Forsberg et al. [2] *	92.1± 3.0	89.7±3.0
Korez et al. [4] *	94.7±4.0	89.0±3.0
Our approach	92.1±1.6	83.72±4.7

Table 2: Comparison of our approach's performance with other state-of-the-art methods on the benchmark dataset of CSI 2014, on the healthy and osteoporotic test sets. (*In these methods, DICE was computed at vertebra level and aggregated; our DICE is computed on entire spine.)

4 Conclusions

In this paper we propose a model-free deep learning based framework for pathological spine segmentation in CT images. Our method uses a pair of fully-convolutional networks that complement one another: The first network provides a coarse localisation of the spine in the form of an attention map, while the second network provides precise high-resolution segmentations. Both these are fused to obtain the final segmentation map of the spine. We evaluate the method on three datasets and obtain promising results indicating applicability in a clinical setting. Our main conclusions are the following: (1) Our approach based on neural networks is robust, generalisable, and precise on fine details as a consequence of its dependence on every voxel and its surrounding, unlike the traditional models that depend on predefined shapes and edges for fine-tuning, (2) our approach successfully segments healthy as well as fractured vertebrae, given both cases are sufficiently represented in the training set, (3) our approach achieves Dice coefficients of above 90% for healthy cases and above 80% for osteoporotic cases on the CSI 2014 dataset. Since our approach is purely data-driven, its performance can be further improved with a larger and more representative dataset. Lastly, (4) we remark that our approach fails to incorporate information pertaining to structural consistency of a spine. This results in a peculiar behaviour where our method fails to segment parts of a vertebra or sometimes entire vertebrae at the start or end of a spine. This can be observed in Figure 5 (top row). We intend to investigate ways in which such global structural regularity can be imposed during the training phase of our networks.

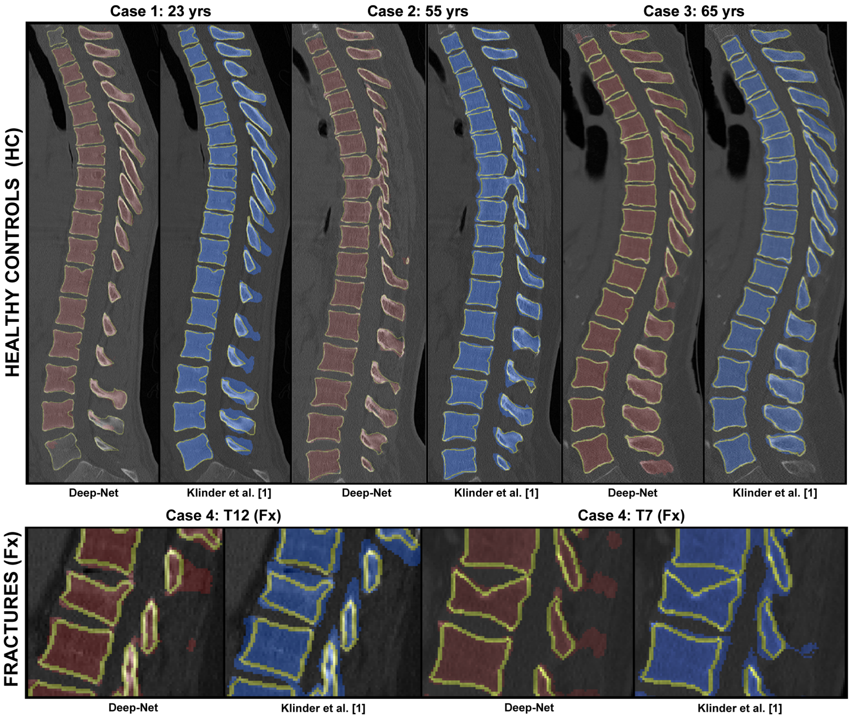

Fig. 5: Comparison of the proposed method (red) and Klinder et *al.* [1] (blue) to the ground truth (yellow contour). **Top row:** Test cases from the Dataset 1. Our method performs better in matching the actual vertebra shape and does not suffer from over-segmentation. However, it does not always provide structurally consistent results; for example, the first and the last vertebra in the left-most image are not fully segmented. **Bottom row**: Fractured cases from the Dataset 2. Our model-free approach is able to capture the unique deformation of the fractured vertebra, opposed to the model-based method.

As part of future work, we plan to employ a graphical model to split the binary segmentation into different vertebrae, thereby labelling the segmentation, based on a conditional random field (CRF)-based approach.

5 Acknowledgements

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

References

1. Tobias Klinder, Jörn Ostermann, Matthias Ehm, Astrid Franz, Reinhard Kneser, and Cristian Lorenz. Automated model-based vertebra detection, identification, and segmentation in ct images. *Medical Image Analysis*, 13(3):471–482, 2009.
2. Daniel Forsberg. Atlas-based segmentation of the thoracic and lumbar vertebrae. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pages 215–220. Springer, 2015.
3. Samuel Kadoury, Hubert Labelle, and Nikos Paragios. Spine segmentation in medical images using manifold embeddings and higher-order mrfs. *IEEE Transactions on Medical Imaging*, 32(7):1227–1238, 2013.
4. Robert Korez, Bulat Ibragimov, Boštjan Likar, Franjo Pernuš, and Tomaž Vrtovec. Interpolation-based shape-constrained deformable model approach for segmentation of vertebrae from ct spine images. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pages 235–240. Springer, 2015.
5. Juying Huang, Fengzeng Jian, Hao Wu, and Haiyun Li. An improved level set method for vertebra ct image segmentation. *Biomedical engineering online*, 12(1):48, 2013.
6. Meelis Lootus, Timor Kadir, and Andrew Zisserman. Automated radiological grading of spinal mri. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pages 119–130. Springer, 2015.
7. Ben Glocker, Johannes Feulner, Antonio Criminisi, D Haynor, and Ender Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In *MICCAI*, pages 590–598. Springer, 2012.
8. Amin Suzani, Abtin Rasoulian, Alexander Seitel, Sidney Fels, Robert N Rohling, and Purang Abolmaesumi. Deep learning for automatic localization, identification, and segmentation of vertebral bodies in volumetric mr images. In *SPIE Medical Imaging*, pages 941514–941514. International Society for Optics and Photonics, 2015.
9. Hao Chen, Chiyao Shen, Jing Qin, Dong Ni, Lin Shi, Jack CY Cheng, and Pheng-Ann Heng. Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks. In *MICCAI*, pages 515–522. Springer, 2015.
10. Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017.
11. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.
12. Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432. Springer, 2016.
13. Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Int'l Conf. on 3D Vision*, pages 565–571. IEEE, 2016.
14. Jianhua Yao, Joseph E Burns, Daniel Forsberg, Alexander Seitel, Abtin Rasoulian, Purang Abolmaesumi, Kerstin Hammernik, Martin Urschler, Bulat Ibragimov, Robert Korez, et al. A multi-center milestone study of clinical vertebral ct segmentation. *Computerized Medical Imaging and Graphics*, 49:16–28, 2016.

15. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
16. A three-dimensional u-net for ptic cleft detection. `https://github.com/zudi-lin/pse-unet`, 2016.
17. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of the 22nd ACM Int'l Conf. on Multimedia*, pages 675–678. ACM, 2014.