Segmentation, Localization and Identification of Vertebra: A Comparison

Mustermann Max
Computer Vision Lab
TU Wien
Vienna, Austria
e000000@student.tuwien.ac.at

***Abstract*—One step in the automatic spine analysis and vertebrae abnormality detection, is the segmentation of vertebrae in either CT or MRI images. Three methods that perform vertebra segmentation together with additional tasks such as localization and identification are examined in this paper. The first method uses instance segmentation that segments vertebrae individually in an iterative approach. The second method builds on the first and extends the fully convolutional network to a two-stage framework. In the first step, bounding box predictions provide a priori knowledge about the anatomy part depicted the second step. The third method performs adversarial learning with a generator and a descriptor in a multi-task framework to make use of the correlation between the individual tasks. The methods are compared with respect to their different approaches and the input data prerequisites. The architecture types and learning strategies as well as their results are discussed. Finally, the strengths and weaknesses are highlighted.**

***Index Terms*—Vertebra, segmentation, localization, identification, convolutional networks, adversarial learning**

I. Introduction

The localization and segmentation of individual vertebrae in either Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) scans are a central step in the computer-assisted analysis of the spine [8], [19]. It is used to automatically detect vertebrae abnormalities [2] or fractures [19]. In clinical workflows, chest or abdomen images are automatically examined for spine diseases, even if the scans are not acquired for this purpose [4]. A challenge in the vertebrae analysis is the variance in the anatomy of the spine and the pathological variants [4]. For example, around 10% of the population have one vertebra less or in addition [16]. Combined with the similarity in appearance, it is difficult to detect the accurate number of vertebrae and subsequently identify and segment each vertebra correctly [20]. Different field of views result in varying region of interests and lead to only partially visible vertebrae [10]. In addition, both modalities, MRI and CT scans, introduce challenges such as varying intensities, varying resolutions, uneven grayscales and imaging noise [11], [20]. Examples for intensity variations (a-c), pathological variants (d-i), uneven grayscales (j-l) and size variants (m-o) are shown in Figure 1.

Starting in 2017, deep learning approaches were introduced to handle the tasks of localization, identification and segmentation of vertebrae [1], [3], [5], [9], [12], [17]. Three representative approaches for whole spine images are discussed and compared in this paper. Lessmann et al. [10] describe an iterative instance segmentation framework that outputs the next not segmented vertebra, the anatomical label and a classification of partial or full visibility. This work is extended by Masuzawa et al. [11] to a two- stage framework that applies an adapted version of the instance framework to cervical, thoracic and lumbar regions. Zhang et al. [20] use an adversarial learning approach to train a generator and descriptor to learn global vertebra relationships.The paper is organized as follows: Section 2 describes the methods which are compared in Section 3 with respect to 1) the application and their prerequisites on the input data, 2) network architectures and training strategies and 3) the datasets used and the corresponding results. Finally, strengths and weaknesses are highlighted and a conclusion is drawn in Section 4.

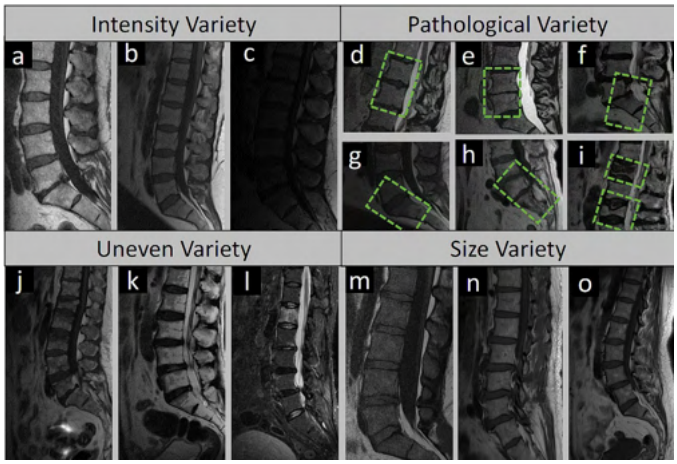


Fig. 1. Challenges depicted in a lumbar spine MRI sagittal slice: intensity variations (a-c), pathological variants (d-e), uneven grayscale (j- l), area and resolution variations (m-o), (taken from [20]).

II. Methods

The first method is an iterative instance segmentation framework by Lessmann et al. [10] that was used as second stage in the framework by Masuzawa et al. [11]. The third method by Zhang et al. [20] uses an adversarial approach. The latter two have been published in 2020 and perform segmentation, localization and identification in one framework.

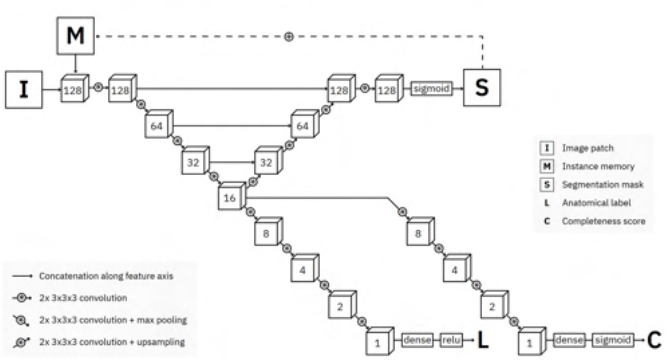


Fig. 2. Schematic network architecture of segmentation network with image patch (I) and instance memory (M) as inputs and segmentation mask (S), visibility classification (S) and anatomical label regression (L) as output (taken from [10]).

A. Instance Segmentation

Lessmann et al. [10] introduce an iterative in- stance segmentation that detects and segments the next not yet segmented vertebrae in an itera- tive manner. The segmentation network is a Fully Convolutional Network (FCN) inspired by the U-Net architecture [13] with an encoder and decoder path connected by skip connections. The architecture is shown in Figure 2. In order to provide the information about already segmented vertebrae of previous iterations to the network, binary labels for each voxel are used as instance memory and fed to the network as auxiliary channel. The iterative approach is realized by a window sliding over the image with the center of the window being adapted after each iteration based on the vertebra parts detected. Two additional branches are attached to the en-coder of the segmentation network. One branch compresses the features further to a single scalar with a sigmoid output layer. This probability value encodes whether the vertebra is partially or fully visible. Incomplete vertebrae are only added to the instance memory but not to the output mask. The other branch is dedicated to predict anatomical labels as a regression output. The labels C1 to L5 are encoded by 1 to 24, 0 is used if the patch does not contain a vertebra. The final vertebrae sequence of labels is determined by taking all regression results into account and calculating the sequence with the highest likelihood.

B. Two-stage Framework

Masuzawa et al. [11] develop a two-stage frame- work that outputs segmentation, localization and identification. The framework architecture is dis-played in Figure 3. The network in Stage 1, called Semantic Segmentation Net, is based on a 3D FCN [14] The aim is to detect bounding boxes for cervical, thoracic and lumbar vertebrae via segmentation. This region information is used as auxiliary information for the second stage. The Iterative Instance Segmentation Net in Stage 2 is based on the iterative segmentation by Lessmann et al. [10]. The segmentation network with instance memory as additional input but without the additional sub-networks is taken to segment the next vertebra not already segmented. Their anatomical labels are counted beginning at the boundary between two subdivisions. In addition, the segmentation result is used to compute the vertebra centroids for localization. The Multi-task Relational Learning Network (MRLN) by Zhang et al. [20] consists of a generator and a discriminator in an adversarial training setup. The framework is visualized in Figure 4. The generator, called Co-Seg-Loc network, produces segmentation mask and localization regression outputs and is inspired by the Spine-GAN by Han et al. [6]. The encoder path uses dilated convolutions to increase the field of view. The Long Short-Term Memory (LSTM) between encoder and decoder learns sequential structures and relationships by memorizing global context from local neighborhoods. The decoder consists of two branches, one for segmentation and the other for localization. In order to learn correlating information, two co-attention modules for Localization-Guided Segmentation Attention (LGSA) and for Segmentation-Guided Localization Attention (SGLA) connect the two branches. The inputs of the discriminator are the ground truth and Co-Seg-Loc network outputs processed to XOR labels that combine the positional relationship of segmentation and localization.

III. Discussion

The following sections compare the papers based on their application and the consequent input preconditions (Section 3.1), the architectures and training strategies used (Section 3.2) and their results and datasets (Section 3.3). The findings are summarized in Table I.

A. Application and Prerequisite

The common task of all three papers is the vertebra segmentation in whole spine images. Lessmann et al. [10] develop their application for CT and MRI scans. Additional outputs are a probability value for the complete visibility of individual vertebra instances and a regression value for the anatomical label. The completeness classification provides the possibility of excluding incomplete vertebra from further processing in automatic spine analysis. Masuzawa et al. [11] and Zhang et al. [20] solve the multi-task of segmentation, identification and localization with different approaches for CT scans and MRIs, respectively. The vertebra localization is not included in the framework by Lessmann et al., but can be performed as post-processing step similar to the work by Masuzawa et al.

All three papers claim to be able to deal with different anatomy parts and image qualities. There-fore, no specific input conditions need to be met for the two-stage framework and the MRLN, expect the correct image modality. Lessmann et al. state that their network might not work for cervical vertebrae and scans with implants close to the spine due to the missing training dataset representation.

B. Network Architectures and Training Strategies

Masuzawa et al. [11] build on the work by Lessmann et al. [10] but extend the framework by a preceding network to predict bounding boxes for thoracic, cervical and lumbar

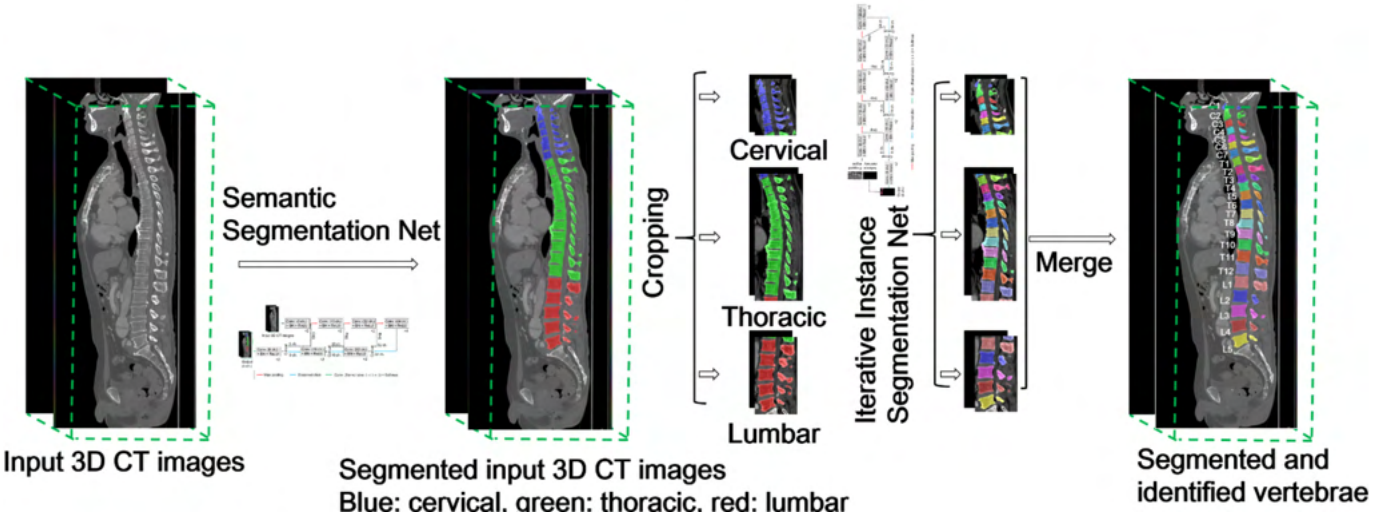


Fig. 3. Two-stage framework with semantic segmentation net and iterative instance segmentation net for vertebra segmentation and localization (taken from [11].

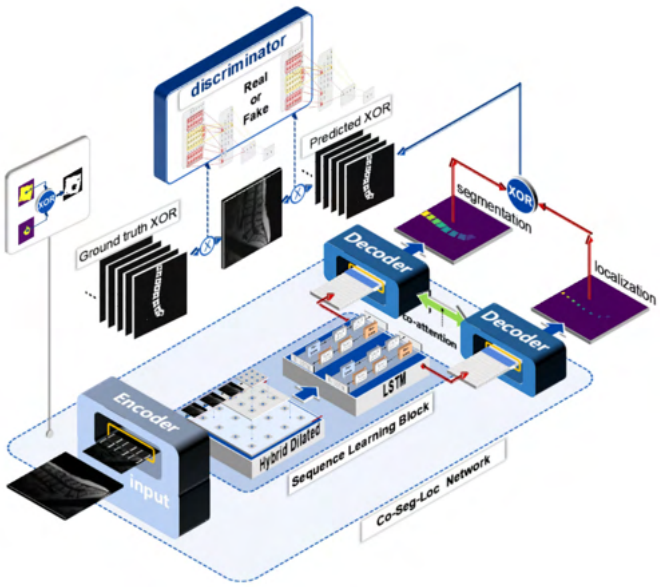


Fig. 4. MLRN architecture with Co-Seg-Loc network as generator for semantic segmentation and localization regression and a discriminator (taken from [20].

regions. Comparing the two iterative instance segmentation networks shows differences in the architecture details. As shown in Figure 5, Masuzawa et al. use 5 layers in the encoder path with filter sizes of 8, 16, 32, 64 and 128 channels, whereas Lessmann et al. use 4 layers with 84 channels from input to decoder connection. The localization step is performed in two ways by Masuzawa et al. and Zhang et al. [20]. Where Masuzawa et al. compute the localization results as post-processing step based on the segmentation result, Zhang et al. predict the localization as regression in a second decoder branch of the Co-Seg-Loc network. The identification step performed by all three frameworks is also handled differently. Zhang et al. do not specify how they compute the identifications, Masuzawa et al. use the segmentation result and a-priori knowledge of anatomy region in a post-processing step and Lessmann et al. predict the anatomical label as regression output.

Different training strategies and loss functions are applied based on the learning tasks. The enumeration is an overview of specified details:

- Lessmann et al. [10] trained their network on a Nvidia Titan X GPU with 12 GB memory for 4-5 days. Their implementation is using the PyTorch framework. They trained two modality-specific networks with a separate CT and MR trainings set. A loss term taking classification, anatomical labeling and segmentation error into account is optimized via Adam with constant learning rate of 0.001 and momentum of 0.99. Data augmentation, such as elastic deformations, Gaussian noise, Gaussian smoothing and cropping is used randomly during training.

- Masuzawa et al. [11] are not providing any information on their software or hardware specifications. However, they use data aug-mentation, especially affine transformations and Gaussian noise during training. The Adam optimizer with a learning rate of 0.001 is used. In the first stage, a boot- strapped cross entropy loss function is used and in the second stage, the Dice loss is optimized.

- Zhang et al. [20] run their Tensorflow based implementation on a Nvidia Titan X GPU. The segmentation network is trained with a multi-task loss and RMSProp

TABLE I
OVERVIEW OF THE COMPARISON OF THE WORK BY LESSMANN ET AL. [10], MASUZAWA ET AL. [11] AND ZHANG ET AL. [20] WITH RESPECT TO APPLICATION AND INPUT PREREQUISITES, NETWORK ARCHITECTURE AND TRAINING STRATEGIES AND DATASETS.

Criteria	Lessmann et al.	Masuzawa et al.	Zhang et al.
Application	Segmentation + Identification + Visibility classification	Segmentation + Localization + Identification	Segmentation + Localization + Identification
Prerequisite	no cervical, implants	none	none
Network architecture	iterative instance segmentation network + branches	two-stage framework with two segmentation networks	Co-Loc-Segm network (generator) and discriminator
Number of networks	1	2	2
Learning tasks	segm. + identif. + class.	loc. + identif.	identif.
Postprocessing tasks	loc. (possible)	loc. + identif.	identif.
Related work	U-Net [13]	FCN [14] & iterative instance segmentation network [10]	Spine-GAN [6]
Optimizer (learning rate)	Adam (0.001)	Adam (0.001)	RMSProp solver & Adam
Framework	Tensorflow	-	PyTorch
Training set	60 CT & 23 MR scans	1035 CT scans	407 subjects
Independent test set	yes for CT & no for MR (three-fold cross validation)	yes (317 CT scans)	no (five-fold cross validation)

solver, whereas the discriminator uses a cross-entropy loss optimized via Adam.

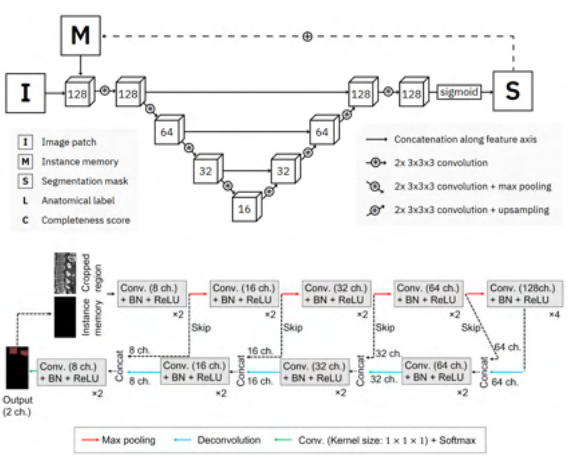


Fig. 5. Comparison of two network architectures: Iterative instance segmentation part of Lessmann et al. (taken from [10], top) and Stage 2 in Masuzawa et al. (taken from [11], bottom).

C. Dataset and Results

Lessmann et al. [10] use 5 different datasets: 15 dedicated thoracolumbar spine CT [18], 15 lumbar spine CT scans [7], a low-dose chest CT with 55 scans [15], 10 lumbar spine CT scans and 23 lumbar spine MR dataset. The split of the training and validation set is performed based on the modality-specific training runs. 3 of the 4 CT datasets were split to training and validation part. The 4th CT dataset was only used for validation. Since the MR dataset only consists of 23 images, three-fold cross validation was used for validation. The Dice co-efficient and the mean Absolute Symmetric Surface Distance (ASSD) are used as error metrics for segmentation. The identification accuracy is defined by the linearly weighted kappa coefficient and the correct labelled vertebrae in percentage. Classification accuracy and average number of false negative and false positives per scan are used for the completeness classification evaluation. Masuzawa et al. [11]use a dataset with 1035 3D CT images for training and independent to that, two datasets with 15 and 302 CT scans [18, 2] for validation. For the segmentation evaluation, ASSD, Hausdorff Distance (HD) and Dice score are used. Euclidean distance and identification rates are evaluation metrics for localization and identification.

Zhang et al. [20] use a dataset of 407 subjects and five-fold cross validation test to evaluate the performance of their approach. The Dice coefficient, the Area Under the ROC Curve (AUC), the localization and identification error are the evaluation metrics. All results are in the same range. Visual results of each method are provided in Figure 6. Since different datasets and error metrics are used to evaluate the performances, the values cannot be compared directly. Table II provides an overview of selected results with Dice score for segmentation, the mean localization error and identification accuracy.

TABLE II
OVERVIEW OF THE SEGMENTATION RESULTS (DICE SCORE IN %), THE MEAN LOCALIZATION ERROR IN MM AND THE MEAN IDENTIFICATION ACCURACY IN %.

Ref.	Dice Score		Localization Error	Id. accuracy	
	CT	MRI		CT	MRI
[10]	94.9 %	94.4 %	-	93 %	100 %
[11]	96.6 %	-	8.3 mm	84 %	-
[20]	-	95.4 %	2.6 mm	-	93.5 %

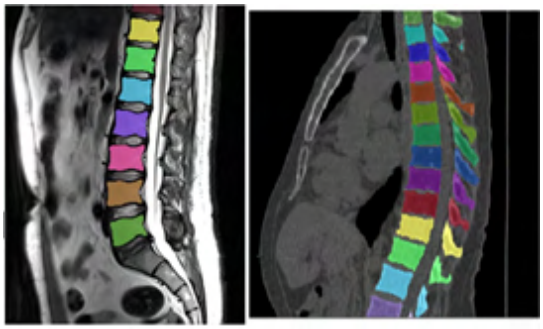
IV. Conclusion

Three approaches to solve the multi-task of vertebrae segmentation, localization and identification were described and compared. The common solved task is vertebrae segmentation in whole spine images of either CT or MRI scans.

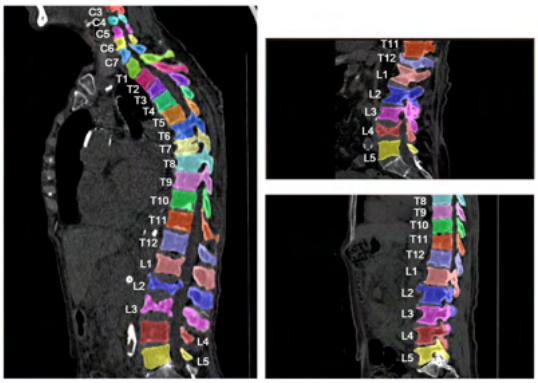
An instance segmentation network, segmenting the next not already processed vertebra in an iterative manner, was extended with two recognition branches to predict vertebrae identification and completeness visibility. The strength of this method is that three tasks are performed with a single network that can be trained end-to-end. Since it was tested for CT and MRI, the architecture works for both modalities without further adaptations. Additionally, the predicted visibility classification is helpful for further analysis. The anatomical labels are based on regression predictions. Therefore, training data annotations are necessary and wrong processing of the vertebrae sequence leads to incorrect anatomical labeling of the image.

In a two-stage framework, the a-priori knowledge about the anatomical region depicted is provided by bounding boxes from an earlier segmentation stage. Thus, assigning anatomical labels and computing the vertebrae centroids is performed via post-processing of the segmentation masks. This is a drawback, since a wrong segmentation mask leads to incorrect localization and identification as well. The two-stage approach has the advantage of having a first stage providing a more detailed knowledge about the anatomy region for the second stage. However, this also means that two networks need to be trained.

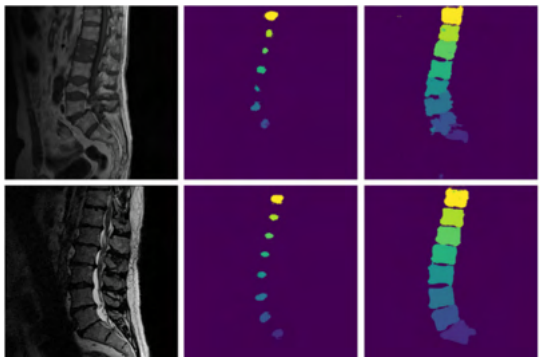
In comparison to that, the MRLN uses an adversarial approach and learns segmentation and localization simultaneously. The framework uses the relationship between the familiar tasks not only for training but also to perform post-processing and correction steps based on the correlated outputs. Although the reported errors are in the same range, a direct comparison of the results is not recommended since different datasets and modalities are used by the authors. Therefore, it would be interesting to see how modified versions of the applications perform on the other modality for the latter two frameworks, or on each other's datasets.



(a) Segmentation for MR (left) and CT (right). [10]



(b) Segmentation and anatomical labels. [11]



(c) Input image (left), localization result (middle) and segmentation result (right). [20]

Fig. 6. Results obtained with the iterative instance segmentation (a), the two-stage framework (b) and the MRLN (c).

References

[1] S. M. M. R. Al Arif, K. Knapp, and G. Slabaugh. Fully automatic cervical vertebrae segmentation framework for X-ray images. Computer Methods and Programs in Biomedicine, 157:95-111, 2018.

[2] I. Ben Ayed, K. Punithakumar, R. Min- has, R. Joshi, and G.J. Garvin. Vertebral body segmentation in MRI via convex relaxation and distribution matching. In Medical Image Computing and Computer-Assisted Intervention MICCAI, Vol. 7510. Springer, Berlin, Heidelberg, 2012.

[3] H.-J. Bae, H. Hyun, Y. Byeon, K. Shin, Y. Cho, Y.J. Song, S. Yi, S.-U. Kuh, J. S. Yeom, and N. Kim. Fully automated 3D segmentation and separation of multiple cervical vertebrae in CT images using a 2D convolutional neural network. Computer Methods and Programs in Biomedicine, 184:105-119, 2020.

[4] C.F. Buckens, Y. van der Graaf, H.M Verkooijen, and et al. Osteoporosis markers on low-dose lung cancer screening chest computed tomography scans predict all-cause mortality. Eu-ropean Radiology, 25:132–139, 2015.

[5] H. Chang, S. Zhao, Y. Chen, and S. Li. Multivertebrae segmentation from arbitrary spine MR images under global view. In Medical image computing and computer-assisted intervention: MICCAI, volume 12266, pages 702–711, 2020.

[6] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li. Spine-GAN: Semantic segmentation of multiple spinal structures. Medical Image Analysis, 50:23 – 35, 2018.

[7] B. Ibragimov, R. Korez, B. Likar, F. Pernus, L. Xing, and T. Vrtovec. Segmentation of pathological structures by landmark-assisted deformable models. IEEE Transactions on Medical Imaging, 36(7):1457–1469, 2017.

[8] A.-A.-Z. Imran, C. Huang, H. Tang, W. Fan, K. Cheung, M. To, Z. Qian, and D. Terzopoulos. Fully-automated analysis of scoliosis from spinal X-Ray images. In IEEE Intl. Symp. on Computer-Based Medical Systems (CBMS), pp. 114–119, 2020.

[9] R. Janssens, G. Zeng, and G. Zheng. Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional

networks. In 2018 IEEE 15th Intl. Sym-posium on Biomedical Imaging (ISBI), pages 893–897, 2018.

[10] N. Lessmann, B. van Ginneken, P. A. de Jong, and I. Igum. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. Medical Image Analysis, 53:142 – 155, 2019.

[11] N. Masuzawa, Y. Kitamura, K. Nakamura, S. Ilzuka, and E. Simo-Serra. Automatic segmentation, localization, and identification of vertebrae in 3D CT images using cascaded convolutional neural networks. In Medical Image Computing and Computer Assisted Intervention MICCAI, volume 12266, pages 681–690. Springer, 2020.

[12] F. Rehman, S. I. Ali Shah, N. Riaz, and S. O. Gilani. A robust scheme of vertebrae segmentation for medical diagnosis. IEEE Access, 7:120387–120398, 2019.

[13] O. Ronneberger and P. Fischer and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention MICCAI, Vol. 9351, Springer, 2015.

[14] H. R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang. An application of cascaded 3D fully convolutional networks for medical image segmentation. Computerized Medical Imaging and Graphics, 66:90 – 99, 2018.

[15] The Natl. Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5):395-409, 2011.

[16] B.J. Tins and B. Balain. Incidence of numerical variants and transitional lumbosacral vertebrae on whole-spine MRI. Insights Imaging, 7:199–203, 2016.

[17] 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, volume 12266, pages 712–722, 2020.

[18] J. Yao, J. E. Burns, D. Forsberg, A. Seitel, A. Rasoulian, P. Abolmaesumi, K. Hammernik, M. Urschler, B. Ibragimov, R. Korez, T. Vrtovec, I. Castro-Mateos, J. M. Pozo, A. F. Frangi, R. M. Summers, and S. Li. A multi-center milestone study of clinical vertebral CT segmentation. Computerized Medical Imaging and Graphics, 49:16–28, 2016.

[19] J. Yao, J.E. Burns, H. Munoz, and R.M. Summers. Detection of vertebral body fractures based on cortical shell unwrapping. Med. Image Comput. Assist. Interv., 15:09–16, 2012.

[20] R. Zhang, X. Xiao, Z. Liu, Y. Li, and S. Li. MRLN: Multi-task relational learning network for MRI vertebral localization, identification, and segmentation. IEEE Journal of Biomedical and Health Informatics, 24(10):2902–2911, 2020.