Deep Learning Framework for Fully Automated Intervertebral Disc Localization and Segmentation from Multi-modality MR Images

Yunhe Gao[1,2(✉)]

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin N.T., Hong Kong
yhgao@link.cuhk.edu.hk
[2] SenseTime Group, Beijing, China

Abstract. Intervertebral discs are joints that lie between vertebrae in the spinal column, which absorb shock between vertebrae during activities. There is a strong correlation between lower back pain and degeneration of intervertebral discs, which may have a great impact on peoples normal life. The precise segmentation of the intervertebral disc is of great significance for the diagnosis of disc degeneration. Currently clinical practice usually manually annotates the volumetric data, which is time-consuming, tedious, needs a lot of expertise and lacks of reproducibility. In this challenge, we developed a fully automated framework that can accurately segment and locate seven intervertebral discs. First, we delicately designed a powerful segmentation network which is a 2D fully convolutional neural network with densely connected atrous spatial pyramid pooling to capture and fuse multi-scale context information. Then we used a localization network and a robust post-process scheme to distinguish different IVD instance. Further more, we proposed a novel training strategy that can make the segmentation network focus on the spine region. The effectiveness of our algorithm is proven in the challenge, we achieved the mean segmentation Dice coefficient of 90.58% and a mean localization error of 0.78 mm.

Keywords: IVD localization · IVD segmentation · Deep learning

1 Introduction

The intervertebral disc is a fibrocartilage disc that connects adjacent vertebrae so that the spine can move within a certain angle. The IVDs have the nature of toughness and elasticity, and can be deformed under pressure, so that the force applied on the IVDs can be evenly distributed into all directions, and ensure the entire surface of vertebral is subjected to the same pressure. IVDs are also the main structure for absorbing shock. When the human body jumps, falls from

G. Zheng et al. (Eds.): CSI 2018, LNCS 11397, pp. 119–129, 2019.
https://doi.org/10.1007/978-3-030-13736-6_10

a high place, and performs other vertical movements, or when the shoulders, back, and waist suddenly load heavy objects, the IVDs can buffer the force by conduction and self-deformation, hence plays the role of protecting the spinal cord and vital organs in the body.

However, with age, excessive activity or overload, it may lead to degeneration of the intervertebral disc, causing lower back pain, numbness of lower limb, nerve injury or even loss of movement, which will seriously affect work ability and life quality. Clinically, medical image analysis is usually the best non-invasive diagnostic method. In order to obtain quantitative parameters, doctors usually manually annotate the IVDs. However, for 3D images, this method is usually tedious, time-consuming, needs a lot of expertise and lack of reproducibility. Therefore, a fully automatic localization and segmentation algorithm of the intervertebral disc can offer visualized 3D reconstructed image and also provide quantitative parameters, which can greatly improve the speed as well as the quality of the diagnosis.

As magnetic resonance imaging has the properties of excellent sensitivity to soft tissue and no radiation, it is widely considered to be the best modality for disc disease diagnosis. Further more, the Dixon method can generate fat only and water only images by combining the in-phase and opposed-phase signal. Making full use of the image information from different modalities can improve the accuracy of the segmentation algorithm. The four modality Dixon sequences are showed in Fig. 1.

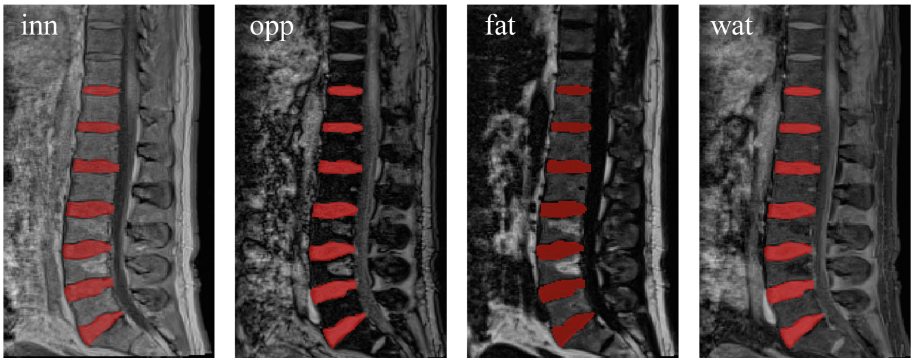

Fig. 1. Examples of multi-modality Dixon sequence, including in-phase, opposed-phase, fat and water from left to right. Each modality has different contrasts for specific components, making full use of multi-modality information can result in better contour segmentation. It should be noted that there are more than seven IVDs in the MR images, but only the lumber IVDs are our objective.

The task of this challenge has two parts, the localization and segmentation of intervertebral disc. The objective of segmentation is to obtain the binary mask of each IVD, i.e. each voxel in the image is classified into the disc category or

non-disc category. The objective of localization is to obtain the coordinates of the centroid of each disc, which is calculated by the morphological center of each IVD mask. The segmentation algorithm affects both segmentation accuracy and localization accuracy, therefore a good segmentation algorithm is a prerequisite.

1.1 Related Work

In early studies, researchers typically used hand-crafted features [4,14] based on image intensity or texture features for IVD localization and segmentation. Graph-based methods are commonly used in the segmentation of vertebrae and discs. For example normalized cut [2] and graph cut algorithm [1] were used for IVDs segmentation in spine MR images. And graphical models [5,12] were used for IVD localization.

As learning-based approaches gain more and more attention in the medical image analysis field, several marginal spacing learning [9] and regression-based methods [3] are proposed for localize IVDs and segment IVDs. However, those methods were limited by the representation capability of the hand-crafted features.

Recently, deep learning methods have revolutionized medical image analysis and computer vision field with its remarkable feature representation capability. For example, Ronneberger et al. [13] proposed U-net for cell segmentation from 2D images and Dou et al. [6] proposed 3D convolutional neural network for 3D liver cancer segmentation. Deep learning methods also improve the performance of IVD localization and segmentation to a brand new level. For example, Li et al. [10] proposed a 3D multi-scale FCN with random modality dropout scheme to better utilize multi-modality information and achieved decent accuracy for IVD localization and segmentation.

1.2 Contribution

We propose a strong and robust deep learning framework for IVDs localization and segmentation from multi-modality MR images. The evaluation results from *MICCAI 2018 Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR Images* demonstrated the effectiveness of our proposed framework. Our main contributions can be summarized as follows:

- We delicately design a 2D fully convolutional network, which only performs downsampling for 2 times, and use densely connected atrous spatial pyramid pooling to capture multi-scale features as well as ensure large enough receptive field. The network consists of three separate pathways for different spatial resolution features, which makes the training of encoder more effective. Further more, a Squeeze-and-Excitation module are used for channel-wise attention. This network is a strong backbone that can be generalized to other medical image segmentation tasks.

- We designed a 3D V-Net based localization network with a robust post-process scheme to classify the seven lumber disc into seven category and distinguish them from other thoracic discs, which makes the whole framework to be fully automated.
- We proposed a novel and intuitive training strategy that can make the segmentation network focus on the spine region while ignore the interference from large and complex backgrounds.
- Our method was evaluated on MICCAI 2018 IVDM3Seg dataset which consists of 16 sets of 3D multi-modality MR images from 8 subjects, and demonstrated superior performance.

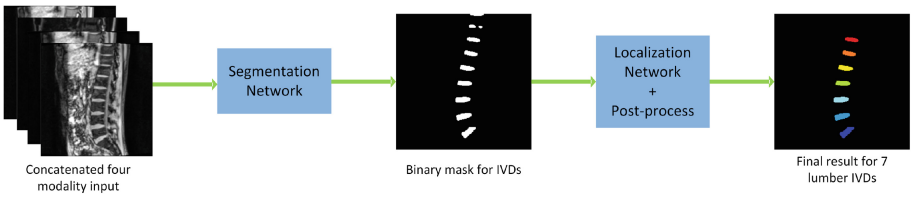

Fig. 2. The pipeline of proposed framework for IVD localization and segmentation. The segmentation first perform binary segmentation to classify each voxel into disc and non-disc region. The localization network and post-process treat each disc instance as a categories and assign label from 1–7 from bottom to top.

2 Methodology

The pipeline of our framework for IVD localization and segmentation are illustrated in Fig. 2. Our localization and segmentation framework mainly consists of two parts: the segmentation network, the localization network and the post-process scheme. The objective of segmentation is to output the binary masks of each IVD, however, because of the similarity between thoracic discs and lumber discs, the segmentation network will predict more than 7 IVD masks, though only 7 lumber discs have annotation. To obtain the final result and achieve the purpose of fully automation, we designed a V-Net based localization network which treats each IVD as an instance, i.e. performs 7 class segmentation, and then used a post-processing method to increase the robustness of the localization network.

2.1 Segmentation Network

In recent years, convolutional neural networks have revolutionized the field of computer vision and medical image analysis. 2D CNNs based methods have made great progress on medical images compared to traditional methods. Recently, 3D CNNs [6,11] are explored as they can capture volumetric contextual information

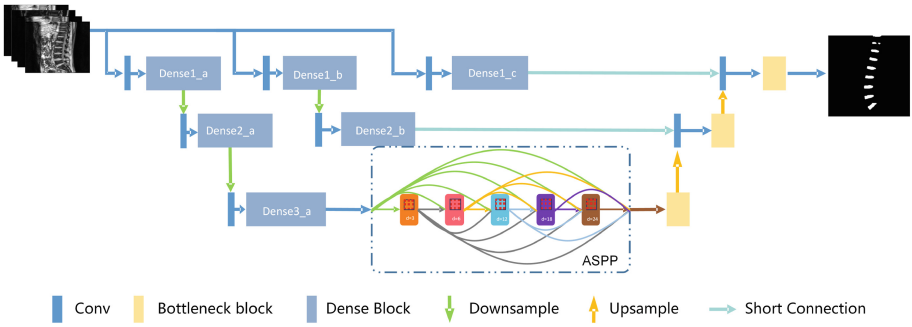

Fig. 3. The proposed 2D fully convolutional segmentation network, it takes the 4-channel concatenated multi-modality image as input, and outputs the binary mask of IVDs.

and have better representation capability. However, 3D CNNs essentially have a disadvantage compared to 2D CNNs, they have a greater demand for data, as 3D CNNs treat a volumetric image as a single sample while 2D CNNs treat each slice as a single sample. There are only 16 samples in the training set, therefore, we think the 2D network is more suitable for this task.

U-net [13] is one of the most successful 2D convolutional neural networks in medical image analysis, many previous deep learning methods are modified based on it. U-net has a symmetric Encoder-Decoder structure, the encoder encodes multi-scale information into feature maps by four downsamplings. The decoder then reconstruct spatial resolution from high-level feature maps by upsampling or deconvolution, while high-resolution features are also concatenated by short connection from encoder to assist reconstruction. However, this structure has three inherent defect for semantic segmentation. First, too many times of down-sampling leads to the loss of detail information, although the high-resolution feature maps are used in the reconstruction process, but this low-level feature concatenate and feature fusion can only slightly alleviate the problem. Second, UNet captures multi-scale features by downsapling, which results in capturing only fixed and limited scales of features, making it difficult to represent complex and variable anatomical structures. Third, during the gradient back propagation in the training phase, the encoder will receive two gradient signals from different resolutions, one is the low resolution gradient signal from below, and the other is the same resolution gradient signal from the shortcut connection. It cannot be guaranteed that the two path have the same magnitude of gradient signal because the number of convolution layers on different paths is quite different. The mixing of these two signals in the training process will affect the effectiveness of the encoder training.

To solve the problems mentioned above, we elaborately design our segmentation network, see in Fig. 3. We use a strong backbone network, which is based on DenseNet [8] and uses Squeeze-and-Excitation module [7] as channel-wise attention. For the first problem, reduce the number of downsampling is a intu-

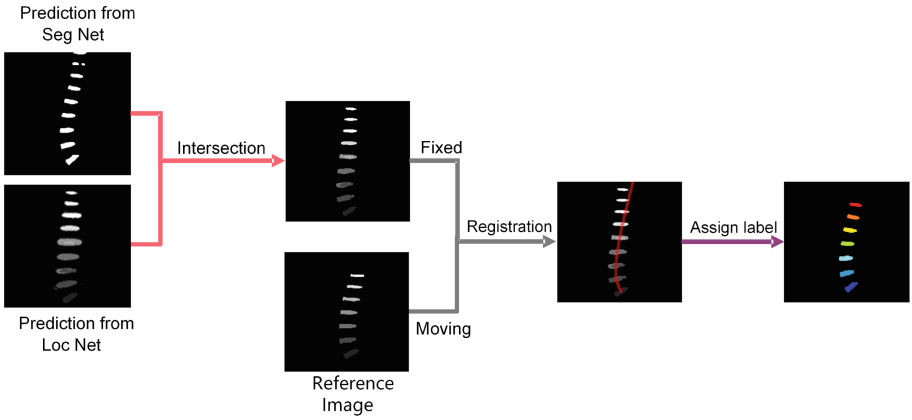

Fig. 4. The pipeline of localization network and the post-process scheme. (Color figure online)

itive solution. In the trade-off between the GPU memory usage and segmentation accuracy, our network only perform two downsamplings, it can effectively reduce the information loss, and improve the segmentation accuracy of the detailed edge region of discs. However, such structure has a disadvantage that the network can only fuse less scales of features, moreover, the receptive field of convolution kernel become smaller, which makes it difficult to capture more global and high-level features. Therefore, we further use densely connected dilated convolution to solve this problem, or use another name, densely connected atrous spatial pyramid pooling (ASPP). Compared with the serial connected or parallel connected [15] counterpart, densely connected ASPP combines arbitrary scales of features, which can be adjusted by dilation rate, and better feature reuse. In our model, we use the dilation rate of 3, 6, 12, 18 and 24. For the third problem, inspired by some works on multi-task learning, we design three separate paths to handle different resolution signal, i.e. treat each resolution signal path as a single task. This approach can train each path more effectively without interfering with each other.

2.2 Localization Network and Post-process

Although the segmentation network is trained only with 7 lumbar discs annotation, the network predicts more than 7 IVDs because of the similar anatomy pattern of thoracic disc and the lumbar disc. We design a localization network and a post-process scheme to handle the output of the segmentation network, and fully automatically get the target mask of 7 lumbar discs. The structure is shown in Fig. 4. The localization network has a V-Net structure, which is a 3D fully convolutional neural network with residual connection. The ground truth annotation of the localization network is obtained by marking the mask of the 7 IVDs in the original annotation from 1 to 7 from bottom to top, that is, the

localization network output 8 channels score map, including seven IVDs and one background.

Then the prediction from localization network and prediction from segmentation network are intersected together. Due to the similar appearance of the IVDs, the predicted mask from localization network in the often have misclassified areas in the upper part of images, but the segmentation of the bottom disc is always right. We then use a reference image from training set as moving image to be registered to the predicted mask, and fit the centerline of spine from the centroid of each disc in the registered mask, i.e. the red line shown in Fig. 4. At last, we calculate the connected area of the predicted mask. Only the connected area that intersects with the fitted centerline of the spine is retained. The other connected regions are set as background. Then, the reserved connected region is assigned with label from 1 to 7 from bottom to top.

This localization and post-processing strategy can greatly improve the robustness of the framework, even if there are some misclassified outliers in the segmentation network, it will not affect the identify of IVDs.

2.3 Training Strategy

To further improve the performance of the segmentation network, we made a natural assumption.

Assumption. Only the spine part of the entire input image is useful for IVD segmentation, while region outside the spine only acts as a useless background, which will reduce the accuracy of the segmentation performance.

We first train a UNet to predict the spine area, where the label was generated by calculating the convex hull of the annotation of discs after several dilation operations. When training the segmentation network, the predicted mask from the UNet was used, and we ignore the loss outside the spine region. In the inference phase, the spine region is also predicted, and all the region outside the spine is set as background in the output of segmentation network.

2.4 Loss Function

When training segmentation network, focal loss was used for better focus on hard samples, i.e. the boundary region of IVDs, and the formula is as follow:

$$L(p) = -\alpha(1-p)^{\gamma} log(p), \tag{1}$$

Since the use of focal loss may cause instability problems when training, we first train several epochs using cross entropy loss, then use focal loss.

3 Experiments

3.1 Dataset and Data Augmentation

We evaluated our proposed method on the dataset from MICCAI 2018 IVDM3Seg Challenge using both cross validation on the training data and independent test data on the on-site challenge, where training data consists of 16

sets of 3D multi-modality MR images from 8 subjects, and test data consists of 8 sets of 3D multi-modality MR images from 4 subjects. Each subject was scanned with a 1.5-Tesla MRI scanner of Siemens using Dixon protocol. The voxel spacing of each image is 2 mm × 1.25 mm × 1.25 mm. For the data augmentation, we use 3D deformation, random scale, random noise, and random crop.

3.2 Evaluation Metrics

Dice overlap coefficient measures the percentage of correctly segmented voxels. Dice is computed by

$$Dice(A, B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \times 100\%, \tag{2}$$

where A is the sets of foreground voxels in the ground-truth data and B is the corresponding sets of foreground voxels in the segmentation result, respectively.

Average absolute distance (ASD) is a metric measures the average absolute distance from the ground truth disc surface and the segmented surface. Smaller average absolute distance means better segmentation accuracy.

Localization distance R is computed by

$$R = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2}, \tag{3}$$

where Δx, Δy, Δz is the absolute difference between the identified IVD center and the ground truth IVD center calculated from the ground truth segmentation in X, Y and Z axis. Smaller localization distance means better segmentation accuracy.

3.3 Results of MICCAI 2018 and Training Set Cross Validation

The evaluation result of the on-site challenge of MICCAI 2018 IVDM3Seg are listed in Tables 1, 2 and 3. Our method demonstrated good performance and strong robustness. Since the test data are not available to us, the segmentation result are visualized using training set cross validation, see in Fig. 5.

Table 1. Dice overlap coefficient of independent test set in on-site challenge.

Dice	Disc_01	Disc_02	Disc_03	Disc_04	Disc_05	Disc_06	Disc_07
Test_01	0.888	0.904	0.927	0.911	0.896	0.890	0.868
Test_02	0.908	0.934	0.940	0.944	0.930	0.923	0.925
Test_03	0.894	0.896	0.900	0.866	0.896	0.818	0.884
Test_04	0.918	0.938	0.938	0.913	0.909	0.885	0.925
Test_05	0.897	0.911	0.917	0.918	0.869	0.896	0.926
Test_06	0.865	0.914	0.929	0.910	0.898	0.898	0.892
Test_07	0.904	0.931	0.931	0.914	0.904	0.887	0.863
Test_08	0.904	0.928	0.922	0.923	0.910	0.907	0.889

Table 2. Average absolute distance of independent test set in on-site challenge.

ASD(mm)	Disc_01	Disc_02	Disc_03	Disc_04	Disc_05	Disc_06	Disc_07
Test_01	0.73	0.67	0.49	0.58	0.63	0.61	0.68
Test_02	0.54	0.47	0.44	0.37	0.44	0.41	0.41
Test_03	0.72	0.82	0.74	0.87	0.57	0.90	0.54
Test_04	0.55	0.48	0.42	0.50	0.48	0.51	0.29
Test_05	0.58	0.70	0.69	0.61	0.97	0.68	0.37
Test_06	0.85	0.66	0.57	0.75	0.78	0.69	0.55
Test_07	0.64	0.53	0.52	0.64	0.62	0.65	0.65
Test_08	0.72	0.63	0.63	0.68	0.62	0.57	0.58

Table 3. Average absolute distance of independent test set in on-site challenge.

Localization(mm)	Disc_01	Disc_02	Disc_03	Disc_04	Disc_05	Disc_06	Disc_07
Test_01	0.44	1.42	0.53	0.78	1.37	0.98	1.08
Test_02	0.38	0.73	0.27	0.47	0.73	0.41	0.18
Test_03	0.64	0.33	1.36	2.11	0.95	1.27	0.46
Test_04	0.83	0.53	0.63	1.35	0.50	1.20	0.08
Test_05	0.80	0.34	0.13	1.04	0.93	1.04	0.49
Test_06	1.23	0.47	0.43	0.12	0.64	1.09	0.66
Test_07	0.60	1.12	0.98	1.21	1.17	0.63	1.07
Test_08	0.28	1.30	0.77	0.92	1.03	0.35	0.60

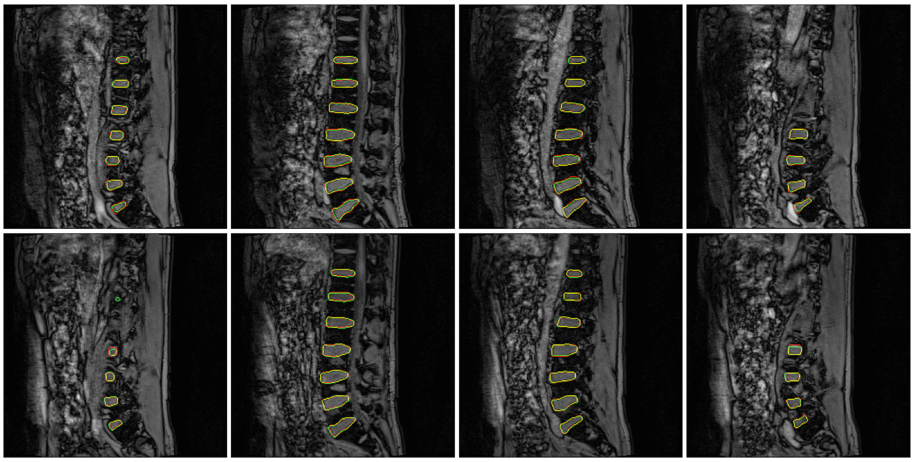

Fig. 5. Visualization of one subject in training set cross validation, the green line is the prediction of our approach, read line is ground truth and yellow line is the intersection. These images are from subject 3, the images in the first row is obtained in the first phase, while the second row is obtained in the second phase. (Color figure online)

4 Conclusion

In this paper, we present our novel and robust IVD segmentation and localization framework from multi-modality MR images, which achieve state-of-the-art performance. The delicately designed segmentation network can preserve the detailed information as much as possible by reducing the number of downsamplings, and at the same time, using densely connected atrous spatial pyramid pooling to capture and fuse multi-scale information as well as reserve large enough receptive field, which can greatly enhance the feature representation ability of the network. We also design three separate paths to handle different resolution signal to train each path more effectively. A new training strategy is also proposed to prevent the segmentation network from interfered by the large complex background. Furthermore, we propose a localization network with robust post-process scheme to distinguish thoracic discs and lumber discs. The result of MICCAI 2018 challenge on IVD localization and segmentation demonstrated the effectiveness of our proposed method.

References

1. Ben Ayed, I., Punithakumar, K., Garvin, G., Romano, W., Li, S.: Graph cuts with invariant object-interaction priors: application to intervertebral disc segmentation. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 221–232. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22092-0_19
2. Carballido-Gamio, J., Belongie, S.J., Majumdar, S.: Normalized cuts in 3-D for spinal mri segmentation. IEEE Trans. Med. Imaging **23**(1), 36–44 (2004)
3. Chen, C., et al.: Localization and segmentation of 3D intervertebral discs in mr images by data driven estimation. IEEE Trans. Med. Imaging **34**(8), 1719–1729 (2015)
4. Chevrefils, C., Cheriet, F., Aubin, C.É., Grimard, G.: Texture analysis for automatic segmentation of intervertebral disks of scoliotic spines from mr images. IEEE Trans. Inf. Technol. Biomed. **13**(4), 608–620 (2009)
5. Corso, J.J., Alomari, R.S., Chaudhary, V.: Lumbar disc localization and labeling with a probabilistic model on both pixel and object features. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008. LNCS, vol. 5241, pp. 202–210. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85988-8_25
6. Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.-A.: 3D deeply supervised network for automatic liver segmentation from CT volumes. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 149–157. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_18
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 (2017)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
9. Kelm, B.M., et al.: Spine detection in CT and MR using iterated marginal space learning. Med. Image Anal. **17**(8), 1283–1292 (2013)
10. Li, X., et al.: 3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images. Med. Image Anal. **45**, 41–54 (2018)

11. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
12. Raja'S, A., Corso, J.J., Chaudhary, V.: Labeling of lumbar discs using both pixel- and object-level features with a two-level probabilistic model. IEEE Trans. Med. Imaging **30**(1), 1–10 (2011)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Schmidt, S., et al.: Spine detection and labeling using a parts-based graphical model. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 122–133. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73273-0_11
15. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890 (2017)