



Automated Segmentation of Intervertebral Disc Using Fully Dilated Separable Deep Neural Networks

Huan Wang, Ran Gu, and Zhongyu Li^(✉)

School of Mechanical and Electrical Engineering,
University of Electronic Science and Technology of China, Chengdu, China
zhongyu.emerald@gmail.com

Abstract. Accurate segmentation of intervertebral discs is a critical task in clinical diagnosis and treatment. Despite recent progress in applying deep learning to the segmentation of multiple natural image scenarios, addressing of the intervertebral disc segmentation with a small-sized training set are still challenging problems. In this paper, a new framework with fully dilated separable convolution (FDS-CNN) is proposed for the automated segmentation of the intervertebral disc using a small-sized training set. Firstly, a fully dilated separable convolutional network is designed to effectively prevent the loss of context information by reducing the number of down-sampling. Secondly, a multi-modality data fusion and augmentation strategy are proposed, which can increase the number of samples, as well as make full use of multi-modality image data. Experimental results validate the proposed framework in the MICCAI 2018 Challenge on Automatic Intervertebral Disc Localization and Segmentation from 3D Multi-modality MR Images, demonstrating excellent performance in comparison with other related segmentation methods.

Keywords: Intervertebral disc · Dilated separable convolution · Semantic segmentation · Multi-modality data fusion

1 Introduction

Disc degeneration is likely to cause various back problems, where accurate segmentation of intervertebral discs (IVDs) from MR images is a critical task in clinical diagnosis and treatment [1]. Recent advances of deep learning techniques have greatly facilitated the segmentation of MR images. Given a MR image (either 2D or 3D), deep learning systems can automatically localize and segment all related lesions end-to-end without user intervention. However, for the segmentation of intervertebral discs, the large range of IVD shapes and the limited

H. Wang and R. Gu—Equal contribution.

number of available datasets pose significant challenges in practical applications, e.g., MICCAI 2018 Challenge on Automatic Intervertebral Disc Localization and Segmentation. Especially, for this challenge, the IVD shapes are dramatically different even in the same type of IVDs, where the available small-sized dataset (only includes 64 3D MR images) are hard to support the training of deep segmentation model. Therefore, new methods need to be developed to address the above challenges for the IVD segmentation.

Early studies on IVD segmentation [2–4] have been done by manually extracted features, where these hand-crafted features are dependent on expert knowledge that can be subjective and unreliable. Recently, with the development of deep learning, many effective methods have been proposed in the field of image segmentation. Lin et al. proposed RefineNet [5], which explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. Zhao et al. [6] proposed PSPNet, which uses the pyramid pooling module to obtain multi-scale features. Wang et al. [7] proposed HDC to reduce the gridding issue caused by the standard dilated convolution operation with a simple and effective method. In particular, Deeplabv3 and Deeplabv3+ proposed by Chen et al. [8,9] achieved a better performance on the PASCAL VOC 2012 semantic image segmentation dataset by using spatial pyramid pooling and dilated convolution.

Unlike natural images, medical images usually lack sufficient annotations to differentiate images or pixels from multi-modality imaging devices [20]. In practice, it is difficult to collect a large number of annotated samples to train the segmentation model. Because of the problem, the above methods have difficulty in adapting to medical images. Recently, researchers have proposed multiple methods focused on the medical image segmentation. For example, Chen et al. [10] proposed a 3D full convolutional network (FCN) for IVD localization and segmentation. Li et al. [11] proposed a multi-scale and modality dropout-learning framework to segment IVDs from four modality MR images. Zeng et al. [12] proposed a deeply supervised multi-scale fully convolutional network, which uses a multi-scale deeply supervised method to automatically segment and locate IVDs and using transfer learning to improve the performance of the deep model. Liao et al. [13] proposed a multi-task 3D FCN combined with a bidirectional recurrent neural network to automatically segment vertebrae from the CT images. In addition, Zeng et al. [14] proposed a deeply supervised 3D fully convolutional network to segment the proximal femur in 3D MR images. However, the problem of losing a lot of contextual information is still well unsolved due to the excessive use of down-sampling. Meanwhile, with the number of network layers increases, the parameters will also increase dramatically, which can increase the computational complexity of the whole network.

Taking the above problems into account, this paper proposes a new framework with fully dilated separable convolution (FDS-CNN) for the automatic segmentation of IVDs, using small-sized training set from multi-modality MR images. Firstly, we design a fully dilated separable convolution network that replaces all standard convolutions with dilated separable convolution, and

prevents the loss of contextual information by reducing the number of down-sampling. At the same time, in the case of ensuring the segmentation performance, the network parameters can be effectively reduced. Subsequently, to make full use of the characteristics of multi-modality data, we propose a multi-modality data fusion and augmentation strategy, which can increase the number of samples in a simple and effective manner, improving the generalization performance of the network. Finally, by drawing on the idea of the attention model [15], we use pre-processing networks to pre-segment the spine and make the network more focused on the places of interest.

This paper is organized as follows. In Sect. 2, we present the proposed framework in detail. Then, our framework is evaluated using the MICCAI 2018 IVD Segmentation Challenge Data Set in Sect. 3. Finally, Sect. 4 draws the conclusion and discusses future works.

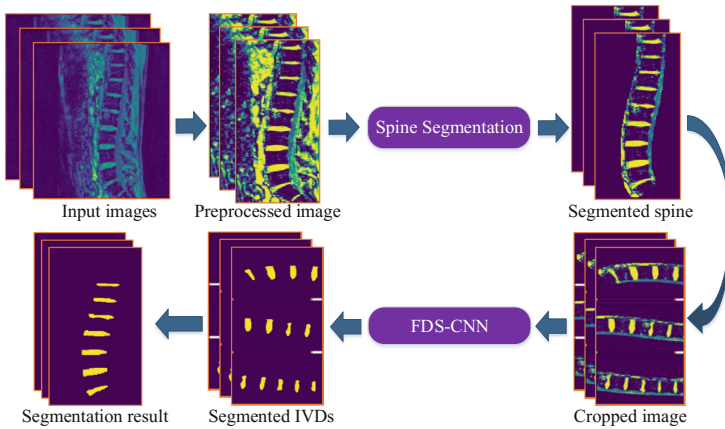


Fig. 1. Overview of the fully automated intervertebral disc segmentation framework.

2 Methodology

Figure 1 presents the overall framework for the automated IVDs segmentation. In order to suppress the complex background interference of multi-modality data, the framework mainly includes two parts, i.e., (1) segmenting the spine out of the original images; (2) segmenting the IVDs using the FDS-CNN. Besides, we propose a multi-modality data fusion and augmentation strategy which can make full use of the characteristics of multi-modality data to effectively increase the number of training samples. Accordingly, in this section, we first present the method for spine segmentation, and then introduce our proposed FDS-CNN structure in detail. Finally, we introduce the multi-modality data fusion and augmentation strategy.

2.1 U-Net for Spine Segmentation

According to Fig. 1, we notice that the original IVD MR image has complicated backgrounds, which may influence the performance of our segmentation model. In order to make the IVD segmentation more focused on the area of interest, we first introduce a pre-processing network to segment the spine from the original image. The network is based on U-net [16] with BN [17] layers after each convolution to speed up network convergence. U-net is a simple and effective semantic segmentation network. It extracts high-level semantic information from images through step-by-step down-sampling, and then restores the size of the image, predicting the results step-by-step through up-sampling and skip connection. Through the pre-processing network, spine images with the area of interests can be obtained. Accordingly, before the segmentation of IVDs, pre-segmentation of spine regions mainly has two advantages, i.e., (1) the subsequent FDS-CNN only need to tackle the area of interest; (2) the computational complexity can be greatly reduced. This idea is similar to the widely used attention model [15] in the field of natural language processing. The model puts more attention on the area of interest to obtain more details of the target and ignore other useless information.

2.2 Convolutional Network with Fully Dilated Separable Convolution

After the pre-segmentation of spine regions, we use the FDS-CNN for the accurate segmentation of IVDs. The FDS-CNN first employs an improved Xception [18] as the encoder network to extract high-level semantic information, extracting multi-scale features based on a spatial pyramid model. Then, it can recover the lost context information using a skip connection. Compared to previous works [6, 8, 9], the propose framework has multiple improvements in the corresponding modules to adapt the IVD segmentation task with only small-sized training set. In particular, our network replaces all convolutions with dilated separable convolutions, which can greatly reduce the number of parameters of the network and effectively extend the field of receptivity. Moreover, our network does not need any pre-training, which can still achieve superior performance with small-sized training set. The following of this section will introduce the implementation details of our network.

Dilated Separable Convolution. The main idea of dilated convolution is to insert “holes” (zeros) between pixels to enlarge the field of convolutional kernels, which enable dense feature extraction in deep CNNs [7]. Dilated convolution allows us to explicitly control the resolution at which feature responses are computed within deep convolutional neural networks [8]. It can effectively expand the field of view in each filter without increasing parameters and computational complexity. We can obtain enough receptive fields through dilated convolution without down-sampling. Therefore, the loss of context information due to down-sampling can be well avoided. Besides, depthwise separable convolution separates the standard convolution into depthwise convolution followed by a

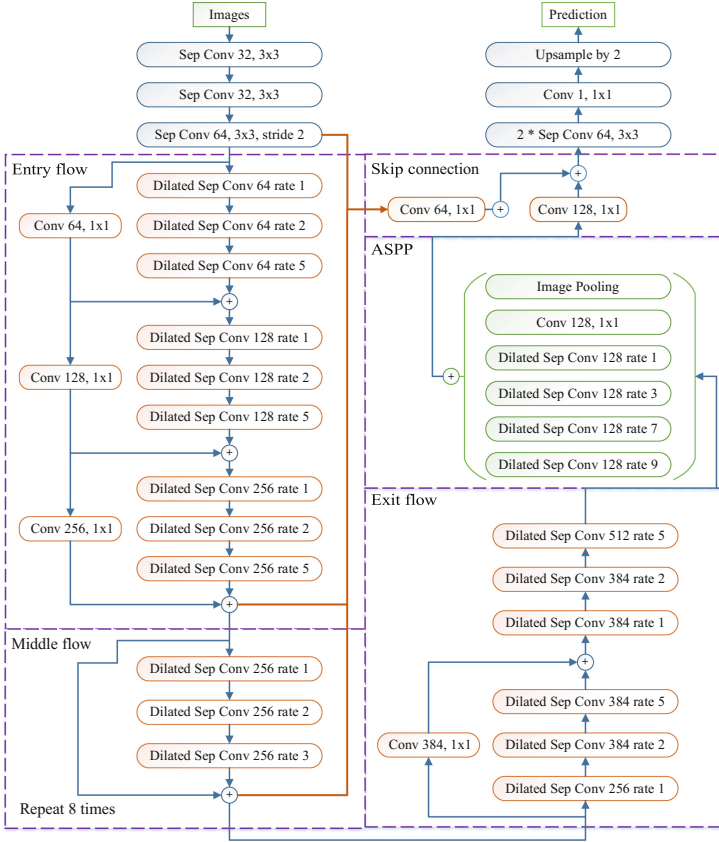


Fig. 2. The structure of the convolutional network with fully dilated separable convolution (FDS-CNN), including modified Xception, ASPP, skip connection, etc.

pointwise convolution. Specifically, the depthwise convolution performs a spatial convolution independently for each input channel, while the pointwise convolution is employed to combine the output from the depthwise convolution [9]. This decomposition can greatly reduce the computational complexity of the model. In our designed FDS-CNN architecture, we use 3×3 depthwise separable convolutions, which can not only have less computation complexity (i.e., 8 to 9 times less) than the standard convolution, but also maintain similar performance as the standard convolution [19]. Our dilated separable convolution combines the depthwise separable convolution and the dilated convolution. The dilated separable convolution embeds the characteristics and inherits the advantages of these two kinds of convolutions. For example, the dilated separable convolution can be treated as a dilated convolution, effectively increasing the receptive field of the network, which also has fewer parameters in comparison with the standard convolution.

Modified Xception. The Xception model [18] has achieved excellent performance in image classification and segmentation tasks. Recently, Chen et al. [9] applied the modified Xception model to address the semantic segmentation and achieve excellent performance. In our solution, we continue to make further three changes to the Xception model and apply it to address the IVD segmentation task. First, we replace all convolutions in the Xception model with dilated separable convolutions and use only one down-sampling in the entire model. Second, in order to further improve the computation efficiency, we reduce the number of all feature maps by half. Third, in order to effectively expanding the receptive field, the dilated separable convolutions in each layer are assigned with different rates. The modified Xception is shown in Fig. 2.

Additionally, we adopt other two strategies to further improve the performance of FDS-CNN, i.e., Atrous Spatial Pyramid Pooling (ASPP) [6,9] and Skip Connection. As shown in Fig. 2, we replace all convolutions in the spatial pyramid structure with dilated separable convolutions, where the rate of dilated separable convolutions in each layer can be modified accordingly. Subsequently, the 1×1 convolutions are applied to three low-level features, which are the output of the third layer convolution, the output of the enter flow and the output of the middle flow, respectively. Then they are concatenation with high-level features. After the concatenation, we apply two 3×3 separable convolutions and one 1×1 convolution to refine the features followed by a simple bilinear upsampling with the factor of 2.

2.3 Multi-modality Data Fusion and Augmentation

For the small-sized training set, the segmentation model is easy to over-fitting. For this problem, a general solution is to increase the number of samples by rotating each image, thereby improving the generalization performance of the model. Although this method can well increase the number of samples, it cannot use the characteristics of multi-modality data itself. Therefore, we develop a new method for multi-modality data fusion and augmentation. According to Fig. 3, from (a-b), it can be seen that multi-modality images have different modalities for the same object (IVD). However, the shape and position of objects inside the image have not changed. Therefore, it is possible to use the feature of multi-modality data to construct new modality. From (e-h), by simply adding the corresponding pixel values from the original two modality images, images with new modality can be obtained. This strategy can not only increase the number of samples, but also fuse different modalities to better represent IVDs. We will verify the performance of the multi-modality data fusion and augmentation strategy in the experimental part.

3 Experiments

3.1 Experimental Setting

We evaluated the proposed framework on the dataset from MICCAI 2018 Challenge of Automatic Intervertebral Disc Localization and Segmentation [21]. The

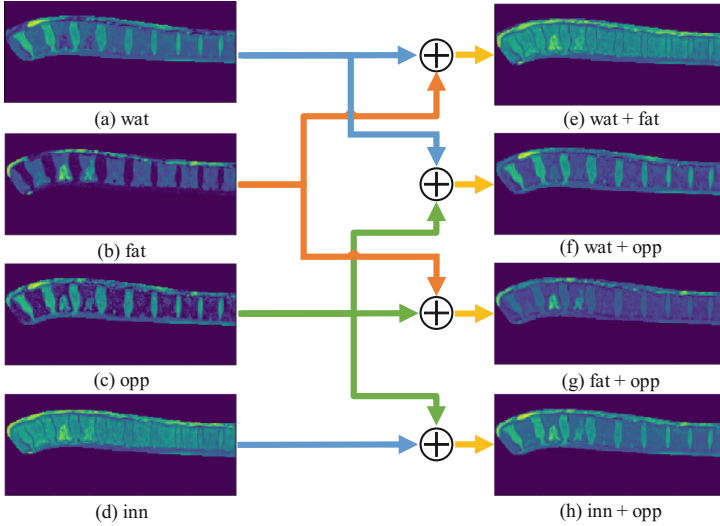


Fig. 3. Schematic diagram of multi-modality data fusion and augmentation strategy.

data set contains 3D images from 8 patients scanned using a 1.5-Tesla MRI scanner at two different times. In addition, each 3D multi-modality MRI data set contains four aligned high-resolution 3D volumes: in-phase (inn), opposed-phase (opp), fat and water (wat) images. There are in total 64 high-resolution 3D MRI volume data. For each IVD, ground truth labels are provided in the form of binary mask with pixel level annotation.

For the data pre-processing, the multi-modality fusion and augmentation strategy discussed above is used to create a variety of sample data, which can significantly increase the number of training samples. Meanwhile, traditional data augmentation strategies such as image rotation are also employed. In order to improve the performance of the model when dealing with blurred imaging samples, we randomly select images before the image input network, i.e., randomly selecting 0%–15% of pixels of the image, assigning random-sized values. With the increasing training batches, each image has different levels of noise, which can increase the diversity of the sample. Besides, each image is normalized using min-max normalization.

For the spine segmentation, in order to reduce the influence of complex backgrounds, we employ a pre-segmentation network to extract spine regions from original images. Spine segmentation network adopt U-net model [16] that has widely applied in image segmentation. After the spine segmentation, we cut the image into 112×128 sub-maps to train the deep neural network. Additionally, the FDS-CNN outputs the predicted image of the same size (112×128) as the training data, where this paper uses the splicing method to restore the predicted image to its original size. In the training of FDS-CNN, we employ the open source architecture from Keras, using the Adam optimization function, where

the learning rate are set as 1E-4, with the batch_size of 16. Our deep neural networks are implemented using Keras on a Linux system with two Nvidia 1080Ti GPUs.

3.2 Evaluation

This paper uses cross-validation to evaluate the performance of the framework. Due to the multi-modality images are scanned at different times, most images from same patients are similar which cannot be set as training and testing data respectively. Therefore, for the cross-validation, the training and testing data will not include the multi-modality images of the same patient. The data set contains in total 16 image data from 8 patients. For each round of validation, 12 image data from 6 patients are selected for training, and the remaining 4 image data are used for testing. In this paper, four groups of cross-validation are performed in each experiment, and dice overlap coefficients are used to evaluate the prediction results of the framework. In the following, we first evaluate the effectiveness of the proposed FDS-CNN, and then verify the performance of the multi-modality data fusion and augmentation strategy.

Table 1. Performance comparison of our network and two benchmarks on the IVD segmentation dataset under different modalities.

	Wat	Fat	Inn	Opp	Mean
Deeplabv3+ [9]	0.8309	0.8124	0.8257	0.8243	0.8235
U-net [16]	0.9107	0.8651	0.9051	0.8992	0.8953
FDS-CNN	0.9111	0.8853	0.9055	0.9062	0.9021

Effectiveness of FDS-CNN. To validate the effectiveness of the proposed FDS-CNN, we compare our approach with 2 benchmark methods: U-net [16] and Deeplabv3+ [9]. For these two benchmarks, as the scanned MRIs are single-channel grayscale images, the input dimensions of networks are modified accordingly. Meanwhile, we reduce the number of channels in the convolutional layers, adding the BatchNormal layer to accelerate the convergence of U-net. All three networks using the same training and augmentation strategies. Table 1 records the dice score of three comparative methods. According to Table 1, the proposed FDS-CNN achieves a mean dice overlap coefficient (MDOC) of 90.21%, where the U-net and Deeplabv3+ only achieve MDOC of 89.53% and 82.35%, respectively. The results demonstrate that the proposed FDS-CNN can achieve better performance in the IVD segmentation task with small-sized training set. Meanwhile, we notice that the accuracy of segmentation achieved by Deeplabv3+ is obviously less than that of U-net. This indicates that Deeplabv3+, which performs well in natural image segmentation tasks, can not well adapt the small-sized medical image data sets. Figure 4 illustrates a randomly selected example with corresponding segmentation results using the proposed FDS-CNN. According

to Fig. 4, our network can achieve the segmentation for IVDs with reasonable results. It is worth pointing out that the accuracy of fat modality in these networks is lower than that of other modalities. This is because the IVDs in the fat modality have low resolution, which can reduce the accuracy of segmentation.

Table 2. The performance of the fused modalities on the trained model.

Wat+Opp	Fat+Opp	Wat+Inn	Inn+Opp	Mean
0.9122	0.9086	0.9121	0.9140	0.9117

Table 3. Results without multi-modality data fusion and augmentation.

Wat	Fat	Inn	Opp	Mean
0.9096	0.8818	0.8984	0.8975	0.8973

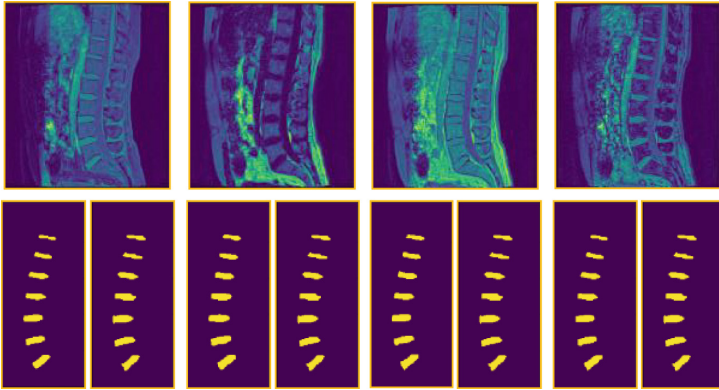


Fig. 4. Examples of segmentation results from the validation data set. From left to right, they are wat, fat, inn, and opp modalities. The second row is their segmentation results (left) and the corresponding ground truth (right).

Validation of Multi-modality Data Fusion and Augmentation. We adopt two protocols to validate the effectiveness of the proposed multi-modality data fusion and augmentation strategy. We first use the new modality to test the segmentation accuracy in the model, and then testing the performance of the model without using multi-modality data fusion and augmentation strategy. According to Table 2, the fused new modalities achieves the MDOC of 91.17%, which is better than the original results, i.e., 90.21% as shown in Table 1. Moreover, the segmentation accuracy of each new modality is also higher than the original modality. In particular, the fat+opp modality is 2.33% higher than the

fat modality result, which can be treated as a preferable solution for the problem of low resolution of the fat modality. This validates that the proposed multi-modality data fusion and augmentation strategy can effectively fuse different features from multiple modalities to improve the accuracy of segmentation. As illustrated in Table 3, the model only achieves MDOC of 89.73% when the multi-modality data fusion and augmentation strategy was not used. Moreover, the segmentation accuracy of each modality is also lower than the results in Table 1. This shows that the multi-modality data fusion and augmentation strategy can provide rich multi-modality data for the network to support the learning of discriminant information, thereby improving the segmentation results of IVDs.

4 Conclusion

In this paper, a new framework with fully dilated separable convolution (FDS-CNN) is proposed for the IVD multi-modality image segmentation with small-sized training set. Compared with other segmentation networks, the proposed FDS-CNN can achieve superior performance in small-sized training set without pre-training. By investigating the information from multi-modality image data, this paper proposes a novel solution for the multi-modality image augmentation, i.e., multi-modality data fusion and augmentation strategy, which can increase the number of samples and improve the performance of the segmentation model. Experiments on MICCAI 2018 IVD Localization and Segmentation Challenge demonstrate the effectiveness and superiority of the proposed framework, in comparison with other state-of-the-arts.

References

1. Luoma, K., Riihimäki, H., Luukkonen, R., Raininko, R., Viikarijuntura, E., Lammien, A.: Low back pain in relation to lumbar disc degeneration. *Spine* **25**(4), 487–492 (2000)
2. 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
3. Law, M.W., Tay, K., Leung, A., Garvin, G.J., Li, S.: Intervertebral disc segmentation in MR images using anisotropic oriented flux. *Med. Image Anal.* **17**(1), 43–61 (2013)
4. Chevrefils, C., Chériet, F., Grimard, G., Aubin, C.-E.: Watershed segmentation of intervertebral disk and spinal canal from MRI images. In: Kamel, M., Campilho, A. (eds.) *ICIAR 2007*. LNCS, vol. 4633, pp. 1017–1027. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74260-9_90
5. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5168–5177 (2017)
6. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *CVPR*, pp. 2881–2890 (2017)

7. Wang, P., et al.: Understanding convolution for semantic segmentation. arXiv preprint, [arXiv: 1702.08502](https://arxiv.org/abs/1702.08502) (2017)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous convolution for semantic image segmentation. arXiv preprint, [arXiv: 1706.05587](https://arxiv.org/abs/1706.05587) (2017)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with Atrous separable convolution for semantic image segmentation. arXiv preprint, [arXiv: 1802.02611](https://arxiv.org/abs/1802.02611). (2018)
10. Chen, H., Dou, Q., Wang, X., Qin, J., Cheng, J.C.Y., Heng, P.A.: 3D fully convolutional networks for intervertebral disc localization and segmentation. In: MICCAI Workshop MIAR, pp. 375–382 (2016)
11. Li, X., Dou, Q., Chen, H., Fu, C.W., Heng, P.A.: Multi-scale and modality dropout learning for intervertebral disc localization and segmentation. In: MICCAI Workshop CSI, pp. 85–91 (2016)
12. Zeng, G., Zheng, G.: DSMS-FCN: a deeply supervised multi-scale fully convolutional network for automatic segmentation of intervertebral disc in 3D MR images. In: Glocker, B., Yao, J., Vrtovec, T., Frangi, A., Zheng, G. (eds.) MSKI 2017. LNCS, vol. 10734, pp. 148–159. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74113-0_13
13. 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 **37**(5), 1266–1275 (2018)
14. Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.-A., Zheng, G.: 3D U-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 274–282. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_32
15. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473) (2014)
16. 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
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456 (2015)
18. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv preprint, [arXiv: 1610.02357](https://arxiv.org/abs/1610.02357) (2017)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint, [arXiv: 1704.04861](https://arxiv.org/abs/1704.04861) (2017)
20. Li, Z., Zhang, X., Müller, H., Zhang, S.: Large-scale retrieval for medical image analytics: a comprehensive review. Med. Image Anal. **43**, 66–84 (2018)
21. IVD3Seg Homepage. <https://ivdm3seg.weebly.com>