Multi-vertebrae segmentation from arbitrary spine MR images under global view

Conference Paper · July 2020

CITATIONS
0

READS
11

5 authors, including:

Heyou Chang
Nanjing Xiaozhuang University
25 PUBLICATIONS **108** CITATIONS
SEE PROFILE

Shen Zhao
Tsinghua University
32 PUBLICATIONS **127** CITATIONS
SEE PROFILE

Shuo Li
The University of Western Ontario
306 PUBLICATIONS **3,587** CITATIONS
SEE PROFILE

Multi-vertebrae segmentation from arbitrary spine MR images under global view

Heyou Chang[1,2,3], Shen Zhao[3], Hao Zheng[1], Yang Chen[2(✉)], and Shuo Li[3(✉)]

[1] School of Information Engineering, Nanjing XiaoZhuang University, Nanjing, Jiangsu, China
[2] School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China
chenyang.list@seu.edu.com
[3] Department of Medical Imaging and Medical Biophysics, Western University, London, Ontario, Canada
slishuo@gmai.com

Abstract. Multi-vertebrae segmentation plays an important role in spine diseases diagnosis and treatment planning. Global spatial dependencies between vertebrae are essential prior information for automatic multi-vertebrae segmentation. However, due to the lack of global information, previous methods have to localize specific vertebrae regions first, then segment and recognize the vertebrae in the region, resulting in a reduction in feature reuse and increase in computation. In this paper, we propose to leverage both global spatial and label information for multi-vertebrae segmentation from arbitrary MR images in one go. Specifically, a spatial graph convolutional network (GCN) is designed to first automatically learn an adjacency matrix and construct a graph on local feature maps, then adopt stacked GCN to capture the global spatial relationships between vertebrae. A label attention network is built to predict the appearance probabilities of all vertebrae using attention mechanism to reduce the ambiguity caused by variant FOV or similar appearances of adjacent vertebrae. The proposed method is trained in an end-to-end manner and evaluated on a challenging dataset of 292 MRI scans with various fields of view, image characteristics and vertebra deformations. The experimental results show that our method achieves high performance (89.28 ± 5.21 of IDR and $85.37 \pm 4.09\%$ of mIoU) from arbitrary input images.

Keywords: Multi-vertebrae segmentation · Global information · Graph convolutional network · Attention network

1 Introduction

Automatic multi-vertebrae segmentation (*i.e.*,partitioning and labeling each of the vertebrae in an input image) from arbitrary magnetic resonance imaging (MRI) is an important step for spinal image analysis and intervention, *e.g.*, spine diseases diagnosis, surgical treatment planning and locating spinal pathologies [1,

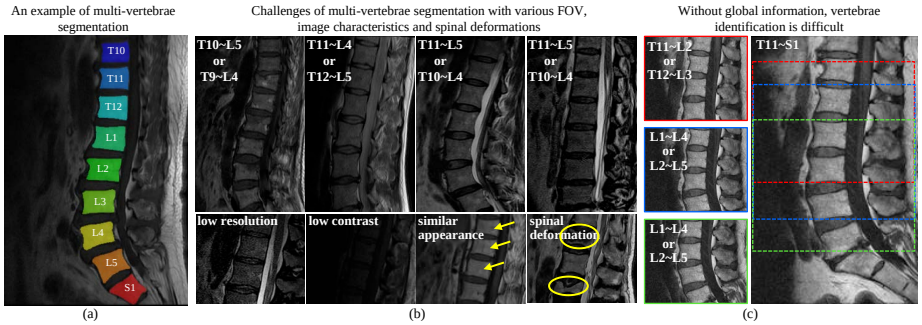

Fig. 1. (a) Each vertebra is partitioned and assigned a unique label (color coded) in multi-vertebrae segmentation. (b) Spine MRI images with various FOV, characteristics and spinal deformations make multi-vertebrae segmentation challenging. (c) Global information plays an important role in multi-vertebrae segmentation.

3, 2]. Precise segmentation is a challenging task due to high topological shape variations, different resolutions and various field of views (FOV), as shown in Fig. 1. Moreover, the similar structures and appearances in close vertebrae and various spinal deformations increase the difficulty of multi-vertebrae segmentation.

The inherent global position of vertebrae is essential prior information for multi-vertebrae segmentation, which has not been fully exploited. Compared to non-vertebra pixels, the pixels belonging to different vertebrae may have more similar appearance. The relative positions of vertebrae are also fixed. Therefore, it is necessary to exploit the relationship between vertebrae in a global view to accomplish multi-vertebrae segmentation effectively. However, most of existing methods with encoder-decoder structure (*i.e.*, FCNs[4], U-net[5]) have a limited receptive field and cannot capture longer-range relationships between vertebrae. For instance, A. Sekuboyina *et al.* [6] employed a multi-layered perception to locate lumbar vertebrae region and adopted 2D U-net to segment and annotate the lumbar vertebrae in the region. R. Janssens *et al.* [7] trained a regression 3D FCNs to find the bounding box of the lumbar region and a 3D U-net to perform segmentation and recognition. N. Lessmann *et al.* [8] first localized and recognized each vertebra in a sliding-window manner, then performed a binary segmentation (spine vs. background) neural network to segment the vertebrae. [11, **?**,10, 12] also proposed other methods such as spine-GAN, cascade amplifier regression network and hierarchical self-calibration detection framework for vertebrae segmentation, detection and/or identification. Due to the lack of global information, these methods consist of multiple phases (vertebra localization and segmentation) with multiple networks to learn, and only work for specific vertebrae segmentation. Taken an arbitrary spine scan as input, there is no end-to-end approach that handles multi-vertebrae segmentation in one go.

Graph convolutional network (GCN)[13] has shown its power for capturing global information of non-grid structure data, and been applied to various com-

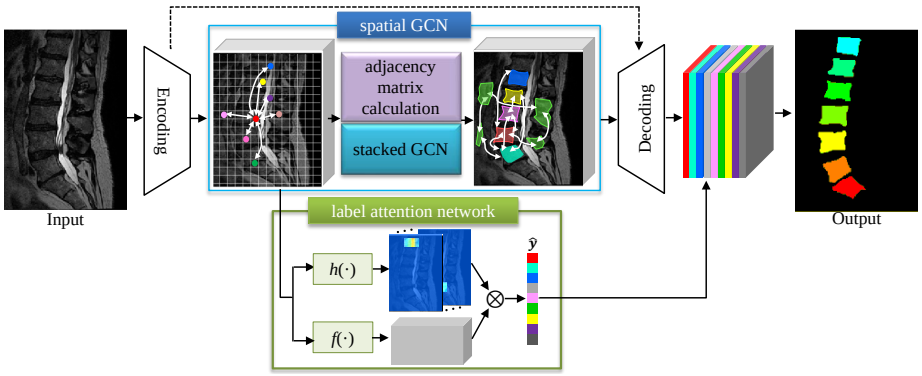

Fig. 2. An overview of the proposed method, which includes (1) a spatial GCN to learn representative features by capturing global spatial information between the vertebrae, (2) a label attention network to reduce the probability of wrong segmentation by exploiting global label information and weighting the output of the decoding network.

puter vision tasks, such as classification [14] and disease prediction [15]. However, these GCN based methods can not be directly applied in vertebrae segmentation because their adjacency matrices are pre-defined and constant, which is improper in segmentation task. Here, we adopt self-attention method to automatically learn the relationships between the pixels and calculate the adjacency matrix based on local feature maps.

In this paper, we propose a novel approach for multi-vertebrae segmentation (MVSeg) with spatial GCN and label attention network, both of which leverage global information to improve segmentation performance. The framework of MVSeg is illustrated in Fig.2. A spatial GCN is introduced to capture *global spatial dependencies* between any two positions of the feature maps. The proposed spatial GCN could effectively model contextual information and generate representative features by constructing a graph on the local feature maps followed by stacked GCN. Then, a label attention network is designed to exploit *global label information* by generating a probability vector of the vertebrae using attention mechanism. The label attention network can reduce the misclassification caused by FOV variety and similarity of adjacent vertebrae by sharing feature learning with decoding network and weighting the output of the decoding network using the probability vector. By exploiting both global spatial and label information, multi-vertebrae segmentation could benefit from long-range information.

Our main contributions are summarized as follows: (1) We propose a novel method for multi-vertebrae segmentation from an arbitrary MRI by leveraging both global spatial and label information. (2) We design a spatial GCN to capture the global spatial information between the vertebrae and a label attention network to exploit the global label information. By modeling global dependencies, the proposed method could effectively reduce the probability of wrong segmentation.

2 Methodology

The proposed MVSeg mainly consists of two parts: (1) Spatial GCN (Section 2.1), which effectively models the inherent dependencies of vertebrae by calculating a adjacency matrix based on local feature maps, and enhances the features' representation capability by globally propagating information using stacked GCN. (2) Label attention network (Section 2.2), which generates a label probability vector through attention mechanism. The probability is used as a weight vector and multiplied with the output of decoding network to produce coherent predictions and reduce the probability of wrong segmentation caused by variant FOV and similar appearance of adjacent vertebrae. An encoding network and a decoding network are also included to learn input representations and recover the spatial resolution, respectively. Skip-connections between them are introduces to enable the two networks to share information.

2.1 Spatial GCN for global spatial information

Spatial GCN captures global spatial relationships of pixels over local feature maps and generates representative features through four carefully designed steps: (1) feature maps reduction to save memory; (2) adjacency matrix calculation to build a graph, (3) stacked GCN to capture longer-range information; and (4) element-wise addition to fuse local and global information. The details of spatial GCN is illustrated in Fig.3.

Given a feature map $\mathbf{X} \in \mathcal{R}^{H\times W\times D}$, we first feed it into a convolution layer with 1×1 filters to reduce the number of feature maps and generate a new feature maps $\hat{\mathbf{X}} \in \mathcal{R}^{H\times W\times d}(d < D)$. Then we reshape $\hat{\mathbf{X}}$ to $\mathcal{R}^{N\times d}$, where $N = H\times W$ is the number of pixels and each row $\hat{\mathbf{x}}_i \in \mathcal{R}^d$ is a feature vector for the i-th pixel. Since each pixel in $\hat{\mathbf{X}}$ corresponds a large patch in the original image and the patches are overlapped, the pixels are related. To calculate the relations between any two pixels, we adopt self-attention method, which outputs a map $\mathbf{S} \in \mathcal{R}^{N\times N}$ by performing a matrix multiplication between $\hat{\mathbf{X}}$ and the transpose of $\hat{\mathbf{X}}$, and applying a softmax layer:

$$\mathbf{S}_{i,j} = \frac{exp(\hat{\mathbf{x}}_i * \hat{\mathbf{x}}_j^T)}{\sum_{j=1}^{N} exp(\hat{\mathbf{x}}_i * \hat{\mathbf{x}}_j^T)} \tag{1}$$

$\mathbf{S}_{i,j}$ represents the value at location (i, j) of $\mathbf{S}$ and measures the similarity of the i-th position and j-th position. The adjacency matrix $\mathbf{A} \in \mathcal{R}^{N\times N}$ is calculated by setting $\mathbf{A}_{i,j} = 0$, if $\mathbf{S}_{i,j} < \tau$; $\mathbf{A}_{i,j} = 1$, otherwise. $\mathbf{A}_{i,j}$ indicates whether i-th and j-th pixels are adjacent or not. τ is a hyper-parameter, which controls the sparseness of $\mathbf{A}$. $\mathbf{A}$ is sparse when τ is large, and vice versa.

Then, GCN is adopted to excavate the global spatial information over the feature map $\hat{\mathbf{X}}$, which can aggregate information from adjacent pixels and output embedding vectors of pixels based on their connections. For a one-layer GCN, the new k-dimensional feature matrix $\mathbf{L}^1 \in \mathcal{R}^{N\times k}$ is computed as

$$\mathbf{L}^1 = f(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\hat{\mathbf{X}}\mathbf{W}_0) \tag{2}$$

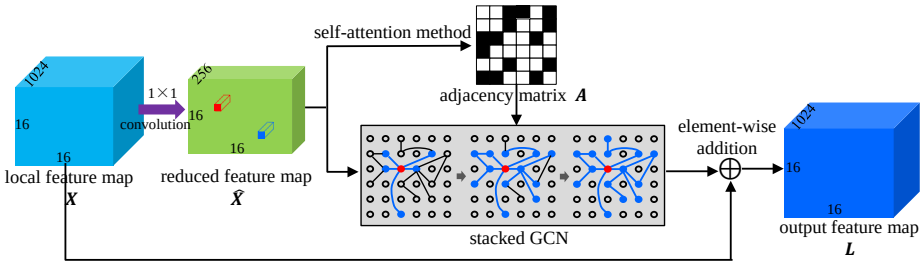

Fig. 3. The details of spatial GCN, which contains four parts: (1) feature maps reduction, (2) adjacency matrix calculation, (3) stacked GCN, and (4) element-wise addition.

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{i,i} = \sum_j \mathbf{A}_{i,j}$. $\mathbf{W}_0 \in \mathcal{R}^{d \times k}$ is a learnable weight matrix, and f is an activation function. The i-th row of $\mathbf{L}^1$ represents the new feature vector of the i-th pixel, which is computed as the weighted sum of its directly adjacent pixels in $\hat{\mathbf{X}}$. GCN with one layer of convolution can only capture information about near neighbors. To incorporate larger neighborhoods information, multiple GCN layers are stacked as follows

$$\mathbf{L}^{i+1} = f(\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{L}^i\mathbf{W}_i) \tag{3}$$

where $\mathbf{W}_i$ is the weight matrix for the i-th layer, and $\mathbf{L}^0 = \hat{\mathbf{X}}$. In the experiment, to feasibly model global information, three layers of GCN are stacked. It can be inferred that each position at $\mathbf{L}^3$ is a weighted sum of features across all positions. The activation function is set as *tanh* at each layer.

The output of the stacked GCN is first reshaped to $H \times W \times k$ and resized to the same size of $\mathbf{X}$ through 1×1 convolution, then added with $\mathbf{X}$ to fuse the original information. It can be seen that the output of the spatial GCN captures both global and local spatial information, which could improve the features' representation ability.

Summarized advantages. The spatial GCN 1) constructs a graph based on local feature maps by regarding each pixel as a node and computing the adjacency matrix by self-attention method, and 2) executes stacked GCN on the graph to acquire larger receptive fields and more representative features.

2.2 Label attention network for global label information

Label attention network generates a label probability vector using attention mechanism and handles the misclassification problem caused by the morphological similarity of adjacent vertebrae and the diversity of FOV. By weighting the output of the decoding network using the label probability vector, the network could produce coherent segmentation results.

The label attention network takes the output $\mathbf{L} \in \mathcal{R}^{H \times W \times D}$ of the spatial GCN as input, and outputs a label probability vector with the dimensionality of C, where each dimension corresponds to one vertebra. In the experiment,

an image contains up to nine vertebrae: T10, T11, T12, L1, L2, L3, L4, L5 and S1. Then C is set to 10 (back ground + nine vertebrae). We first design an attention estimator $h(\cdot)$ to automatically generate a label attention map $\mathbf{B} = h(\mathbf{L})$, where $\mathbf{B} \in \mathcal{R}^{H \times W \times C}$. $h(\cdot)$ consists of 3 convolution layers with 512 kernels of size 1×1, 512 kernels of size 3×3, and C kernels of size 1×1, respectively, followed by batch normalization and ReLu nonlinearity operations except the last convolution layer. Then, $\mathbf{B}$ is normalized per channel by $B_{i,j,c} = \frac{exp(B_{i,j,c})}{\sum_{i,j} exp(B_{i,j,c})}$. Intuitively, the value of $B_{i,j,c}$ should be higher if label c is related to the input image and the image region corresponding to location (i, j) is related to label c. After that, the label attention map $\mathbf{B}$ is used to calculate a weighted feature vector for each label by $\mathbf{f}_c = \sum_{i,j} B_{i,j,c} \mathbf{L}_{i,j,:}$, where $\mathbf{L}_{i,j,:} \in \mathcal{R}^D$. The weighted feature vector $\mathbf{f}_c$ is related to image regions corresponding to label c, which could improve the robustness and accuracy of classification.

To predict the appearance of each label, a linear classifier is learnt for each label by $\hat{y}_c = \mathbf{W}_c \mathbf{f}_c + \mathbf{b}_c$, where $\mathbf{W}_c$ and $\mathbf{b}_c$ are classifier parameters for label c. Considering $\hat{y}_c = \sum_{i,j} B_{i,j,c}(\mathbf{W}_c \mathbf{L}_{i,j,:} + \mathbf{b}_c)$, which can be viewed as performing a multiplication and weighted aggregation at every location of $\mathbf{L}$, the linear classifiers for all labels are modeled as a convolution layer $f(\cdot)$ with C kernels of 1×1 in the implementation. The output of $f(\cdot)$ multiplies with $\mathbf{B}$ element-wisely, and generates a label confidence vector $\hat{\mathbf{y}} = [\hat{y}_1, ..., \hat{y}_C]$ by performing sum-pooling and sigmoid nonlinearity operations. The loss function of the label attention network is defined as $\ell_{classification} = -\sum_{i=1}^{C}(\hat{y}_i * logy_i + (1 - \hat{y}_i) * log(1 - y_i))$, where y_i represents the ground label truth. $y_i = 1$ if and only if the i-th vertebra is associated with image, and 0 otherwise.

Then, the label probability vector $\hat{\mathbf{y}}$ is used as a weight vector and multiplied with the output of the decoding network $\mathbf{S} \in \mathcal{R}^{256 \times 256 \times C}$ by $\hat{\mathbf{S}}_{:,:,c} = \hat{y}_c \mathbf{S}_{:,:,c}$. The cross-entropy loss between the ground segmentation truth and $\hat{\mathbf{S}}$ is adopted as the segmentation loss $\ell_{segmentation}$. Finally, the whole network is trained using a loss function consisting of the segmentation loss and classification loss:

$$\ell = \lambda \cdot \ell_{segmentation} + (1 - \lambda) \cdot \ell_{classification} \tag{4}$$

where λ is a parameter to balance the segmentation and classification loss.

Summarized advantages. The label attention network 1) generates a label probability vector based on multiple linear classifiers and weighted feature vectors using attention mechanism, and 2) multiplies the segmentation output with the probability vector to improve the segmentation performance.

3 Experiments and results

Dataset. The proposed MVSeg is evaluated on a challenging dataset, which consists of 292 arbitrary MR images with various FOV, image characteristics and vertebrae deformations. There are seven kinds of FOV in all, *i.e.*, T10∼S1(63), T11∼S1(105), T12∼S1(64), T10∼L5(19), T10∼L4(12), T11∼L5(15) and T11∼L4 (14). The image number of each kind is listed in the brackets. It can be seen

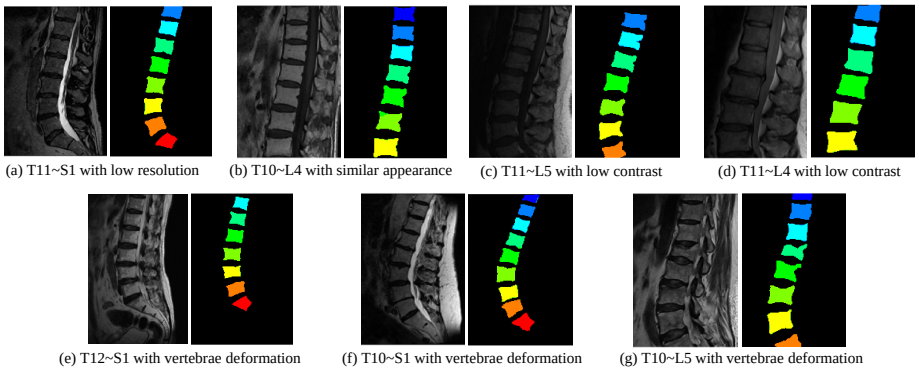

Fig. 4. The proposed MVSeg achieves high multi-vertebrae segmentation performance on a challenging dataset with varied FOV, different image characteristics and vertebrae deformations. Left: input image. Right: segmentation result. The best view in color.

that the distributions of FOV in the dataset are quite different. The images are automatically extracted from each 3D MRI scan and resized to 256×256. The proposed approach is implemented using the Tensorflow backend and trained on Nvidia P100 GPUs with 16 GB memory for 10000 steps with a fixed learning rate of 0.0001.

Evaluation metrics. Standard five-fold cross-validation is adopted for evaluation, and four metrics are used to evaluate the segmentation performance: 1) average precision (AP), which is the percentage of pixels in the total pixels that are segmented correctly in all images; 2) identification rate (IDR), which measures the accuracy of individual vertebra recognition; 3) dice coefficient (Dice), which quantifies the similarity between the segmentation and ground truth of all vertebrae in all images; and 4) mean region intersection over union (mIoU), which is the average IoU of the segmentation and ground truth mask of all vertebrae in all images.

Qualitative demonstration. Fig.4 demonstrates that MVSeg achieves high multi-vertebrae segmentation performance. The partition of each vertebrae is highly precise and the recognition of each vertebra is correct in arbitrary images with varied FOV, image characteristics and vertebrae deformations. For example, (b), (c) and (e) have 7 vertebrae, but their FOV and image characteristics are different.

Quantitative analysis. Table 1 gives the mean accuracy and variance of MVSeg and U-net, which is a well-known segmentation network for biomedical images. Moreover, to verify the effectiveness of the spatial GCN and the label attention network on segmentation accuracy, the performance of MVSeg without spatial GCN ($\text{MVSeg}_{/s}$) and MVSeg without label attention network ($\text{MVSeg}_{/l}$) is also listed in Table 1. It is worth noting that the parameter settings in the encoding and decoding path of all methods are the same. From Table 1, we can see that MVSeg achieves the best performance on all of the four metrics. Compared

Table 1. The mean accuracy and variance (%) of different methods for multi-vertebrae segmentation

	AP	IDR	Dice	mIoU
U-net	97.33±2.08	80.83±11.85	78.14±10.91	76.24±12.17
MVSeg$_{/l}$	98.53±1.35	85.36±5.25	81.93±4.91	80.65±5.94
MVSeg$_{/s}$	98.64±0.24	88.50±2.64	84.80±2.97	82.58±2.42
MVSeg	**98.80±0.49**	**89.28±5.21**	**87.09±4.09**	**85.37±4.09**

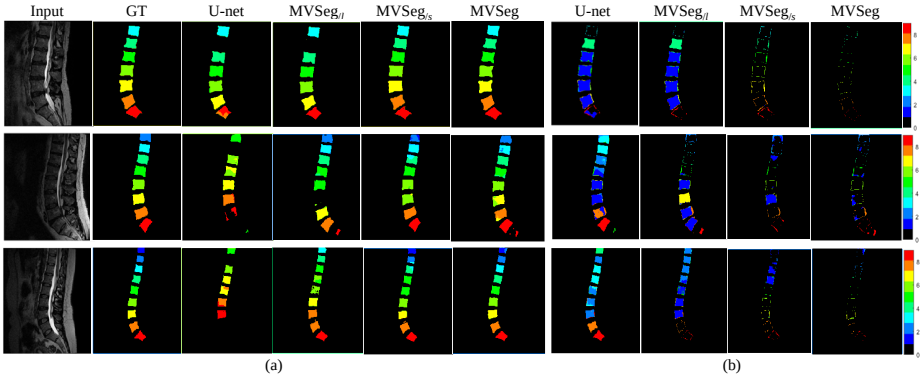

Fig. 5. Some multi-vertebrae segmentation results of different methods. (a) segmentation results, (b) error images. The best view in color.

with U-net, the improvements of MVSeg on AP, IDR, Dice and mIoU are **1.47**%, **8.45**%, **8.95**% and **9.13**%, respectively. Both MVSeg$_{/s}$ and MVSeg$_{/l}$ perform better than U-net with higher mean accuracy and lower variance, which indicate the validity of the two proposed modules in multi-vertebrae segmentation.

Fig. 5 presents visualization of some segmentation results of testing images. The first and second columns represent input image and ground truth (GT), respectively. The third column is the segmentation result of U-net, which has poor precision and accuracy of segmentation. The fourth and fifth columns are the results of MVSeg$_{/l}$ and MVSeg$_{/s}$, respectively. Although their performance is better than U-net, the recognitions of middle vertebrae are mixed with each other and the boundary of some vertebrae is not accurate. By fusing both global position and label information, MVSeg can segment and label the vertebrae more precisely (the last column in Fig. 5(a)). Fig. 5(b) shows the error images between the outputs of different methods and the ground truth.

4 Conclusion

In this paper, we propose a novel method for multi-vertebrae segmentation from arbitrary MRI by incorporating global spatial and label information. In MVSeg, a spatial GCN is constructed to enhance the features' representation ability

by constructing a graph and adopting stacked GCN to capture global position information; and a label attention network is designed to exploit global label information using attention mechanism. The method is trained in an end-to-end manner and verified on a challenging dataset with various FOV, image characteristics and vertebrae deformations. The experimental results demonstrate the effectiveness of MVSeg.

Acknowledgement. This work was supported in part by the National Natural Science Fund of China under Grant 61806098 and 61976118, in part by the State's Key Project of Research and Development Plan under Grant 2017YFA0104302, 2017YFC0109202 and 2017YFC0107900.

References

1. Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S.: Multi-modality vertebra recognition in arbitrary views using 3d deformable hierarchical model. IEEE transactions on medical imaging 34(8), 1676-1693(2015). doi:10.1109/TMI.2015.2392054
2. Han, Z., Wei, B., Leung, S., Nachum, I., Laidley, D., Li, S.: Automated pathogenesis-based diagnosis of lumbar neural foraminal stenosis via deep multiscale multitask learning. Neuroinformatics 16(3-4), 325-337(2018).
3. Liao, H., Mesfin, A., Luo, J.: Joint vertebrae identification and localization in spinal ct images by combining short-and long-range contextual information. IEEE transactions on medical imaging 37(5), 1266-1275(2018). doi:10.1109/TMI.2018.2798293
4. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431-3440. IEEE Press, Boston(2015). doi:10.1109/CVPR.2015.7298965
5. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. LNCS, volume 9351, pp. 234-241. Springer, Munich(2015). doi:10.1007/978-3-319-24574-4_28
6. Sekuboyina, A., Valentinitsch, A., Kirschke, J.S., Menze, B.H.: A localisation-segmentation approach for multi-label annotation of lumbar vertebrae using deep nets. arXiv preprint arXiv:1703.04347(2017)
7. Janssens, R., Zeng, G., Zheng, G.: Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks. In: 15th IEEE International Symposium on Biomedical Imaging. pp.893-897. IEEE Press, Washington(2018). doi:10.1109/ISBI.2018.8363715
8. Lessmann, N., van Ginneken, B., Isgum, I.: Iterative convolutional neural networks for automatic vertebra identification and segmentation in ct images. In: Medical Imaging 2018: Image Processing. vol. 10574, p.1057408. International Society for Optics and Photonics, Huston(2018). doi:10.1117/12.2292731
9. Han, Z., Wei, B., Mercado, A., Leung, S., Li, S.: Spine-GAN: Semantic segmentation of multiple spinal structures. Medical image analysis 50, 23-35(2018)
10. Pang, S., Leung, S., Nachum, I.B., Feng, Q., Li, S.: Direct automated quantitative measurement of spine via cascade amplifier regression network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS, volume 11071, pp.940-948. Springer, Granada(2018). doi:10.1007/978-3-030-00934-2_104

11. He, X., Zhang, H., Landis, M., Sharma, M., Warrington, J., Li, S.: Unsupervised boundary delineation of spinal neural foramina using a multi-feature and adaptive spectral segmentation. Medical image analysis 36, 22-40(2017).
12. Zhao, S., Wu, X., Chen, B., Li, S.: Automatic vertebrae recognition from arbitrary spine mri images by a hierarchical self-calibration detection framework. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS, volume 11767, pp. 316-325. Springer, Shenzhen(2019). doi:10.1007/978-3-030-32251-9_35
13. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in neural information processing systems. pp. 3844-3852. Curran Associates, Barcelona(2016)
14. Zhang, M., Cui, Z., Neumann, M., Chen, Y.: An end-to-end deep learning architecture for graph classification. In: 32th Thirty-Second AAAI Conference on Artificial Intelligence, pp. 4438-4445, AAAI Press, New Orleans(2018)
15. Kazi, A., Shekarforoush, S., Krishna, S.A., Burwinkel,et.al: Graph convolution based attention model for personalized disease prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. LNCS, volume 11767, pp. 122-130. Springer, Shenzhen(2019). doi:10.1007/978-3-030-32251-9_14