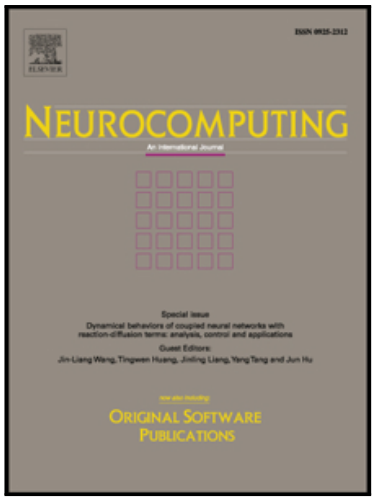
Journal Pre-proof



MMCL-Net: Spinal Disease Diagnosis in Global Mode using Progressive Multi-task Joint Learning

Yanfei Hong, Benzheng Wei, Zhongyi Han, Xiang Li, Yuanjie Zheng, Shuo Li

PII: S0925-2312(20)30300-3
DOI: https://doi.org/10.1016/j.neucom.2020.01.112
Reference: NEUCOM 21970

To appear in: *Neurocomputing*

Received date: 4 April 2019
Revised date: 24 December 2019
Accepted date: 30 January 2020

Please cite this article as: Yanfei Hong, Benzheng Wei, Zhongyi Han, Xiang Li, Yuanjie Zheng, Shuo Li, MMCL-Net: Spinal Disease Diagnosis in Global Mode using Progressive Multi-task Joint Learning, *Neurocomputing* (2020), doi: https://doi.org/10.1016/j.neucom.2020.01.112

Highlights

- Multiple structures of the spine are directly interdependent and influential, and the multi-tasks under a deep convolutional neural network framework can also influence each other.
- Densely dilate ResNet module combat the challenge of size discrepancy and extract significant radiological features.
- Integrate the variational level-set function can enrich the local morphological features of the deep learning model.
- Applied potential features practical can assist other tasks.

MMCL-Net: Spinal Disease Diagnosis in Global Mode using Progressive Multi-task Joint Learning

Yanfei Hong[a,b], Benzheng Wei[a,b,*], Zhongyi Han[c], Xiang Li[a,b], Yuanjie Zheng[d], Shuo Li[e,f]

[a] *College of Science and Technology, Shandong University of Traditional Chinese Medicine, Jinan, SD, China*
[b] *Center for Medical Artificial Intelligence, Shandong University of Traditional Chinese Medicine, Jinan, SD, China*
[c] *School of Software, Shandong University, Jinan, SD, China*
[d] *School of Information Science and Engineering, Shandong Normal University, Jinan, SD, China*
[e] *Digital Image Group (DIG), London, ON, Canada*
[f] *Dept. of Medical Imaging, Western University, London, ON, Canada*

Abstract

Simultaneous detection, segmentation, and classification of multiple spinal structures on MRI is crucial for the early and pathogenesis-based diagnosis of multiple spine diseases in the clinical setting. It is more assistance for radiologists reflections on the disease based on the pathogenesis when the lesion area and its adjacent structures are detected. Obviously, the multiple structures of the spine are directly interdependent and influential, and the multi-tasks under a deep convolutional neural network framework can also influence each other. Multi-task joint optimization in the spinal global mode is a direct outlet to seek the dynamic balance of the above potential correlation. In this paper, we propose a novel end-to-end Multi-task Multi-structure Correlation Learning Network (MMCL-Net) for the detection, segmentation, and classification (normal, slight, marked, and severe) of three types of spine structure: disc, vertebra, and neural foramen simultaneously. And the model is locally optimized to achieve a more stable dynamic equilibrium state. Extensive experiments on T1/T2-weighted MR scans from 200 subjects demonstrate that MMCL-Net achieves high per-

*Corresponding author
Email address: wbz99@sina.com (Benzheng Wei)

formance with mAP of 0.9187, the classification accuracy of 90.67%, and dice coefficient of 90.60%. The experimental results show that the performance of our method is comparable to that of the state-of-the-art methods.

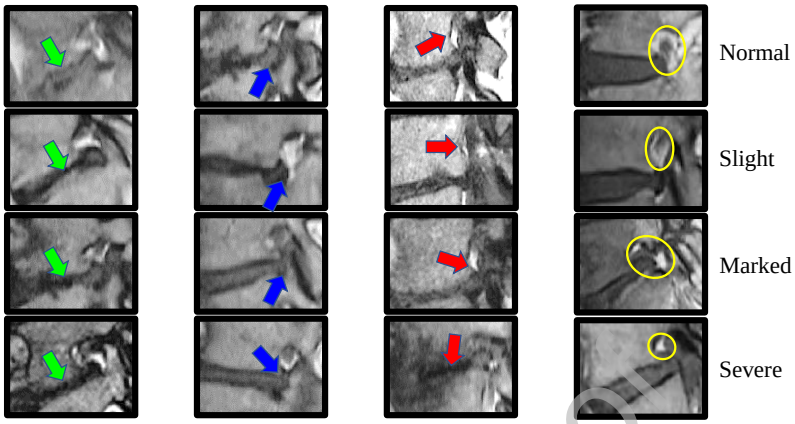
Keywords: densely aggregation, level-set, global optimization, progressive multi-task, multi-structure, medical image

2010 MSC: 00-01, 99-00

1. Introduction

Simultaneous detection, segmentation, and classification of intervertebral disc, vertebrae, and neural foramen is a less trivial study in the early detection and clinical diagnosis of various spinal diseases. Multiple spinal diseases not only have deteriorated the quality of life but have high morbidity worldwide (Rajaee et al., 2012). Lumbar Neural Foraminal Stenosis (LNFS) has attacked about 80% of the elderly population (Lee et al., 2010). In daily radiologist practice, time-consuming and labor-intensive manual readings of MR scans lead to heavy tasks for radiologists, increasing the waiting time of patients and the cost of hospital resources. Therefore, automated analyses of key spinal structures are extremely necessary for radiologists to improve their work efficiency and avoid missing diagnosis, especially in the primary level hospitals.

There is a direct connection between the multiple structures of spine based on the pathogenesis. Simultaneous analysis of intervertebral disc, vertebrae, and lumbar neural foramen is beneficial for the pathogenesis-based diagnosis, which draws crucial pathological links between spinal diseases and its pathogenic factors (Han et al., 2018a). For instance, degeneration of the intervertebral disc and narrow of the intervertebral space lead to decreasing height of the lumbar neural foramen (longitudinal stenosis), and intervertebral disc herniation leads to narrowing width of the neural foramen (lateral narrowing), as shown in Fig. 1(a, b, c). According to statistics, among our 200 clinical patient samples, there are 1633 pathological changes of spine structures, 841 of which has two or more simultaneous lesions. In terms of algorithm research, multi-task has been



(a) Disc Stenosis (b) Disc Herniation (c) Degradation (d) LNFS

Figure 1: Various examples of LNFS caused by different pathogenic factors. (a)-(c) columns showing LNFS caused by disc degeneration, disc herniation, and structure (neural foramen, vertebrae) degradation. Yellow circled areas in (d) column represent instances of four different grades of LNFS according to the Wildermuth qualitative grading system.

used for a variety of visual researches since (Kokkinos, 2017) introduced a low-, medium-, and high-level visual task in the same architecture. It has significant unique ascendency over single scene learning. Specifically, multi-task not only makes related tasks benefit from each other by sharing representation learning but can alleviate the problem of insufficient data to a certain extent on medical images.

No work has so far achieved simultaneous detection, segmentation, and classification of intervertebral disc, vertebrae, and neural foramen due to its numerous difficulties. From the perspective of multiple spine structures, they have the following characteristics: (1) Complicated tasks: each spine has at least 17 target structures, simultaneous localization, segmentation, and diagnosis of all structures are extremely difficult than individual tasks (Han et al., 2018a). (2) Complex structure: the multiple spine structures are interrelated and dependent, and there are large dimensional distinction among the different structures. Besides, normal and abnormal structures have extremely high

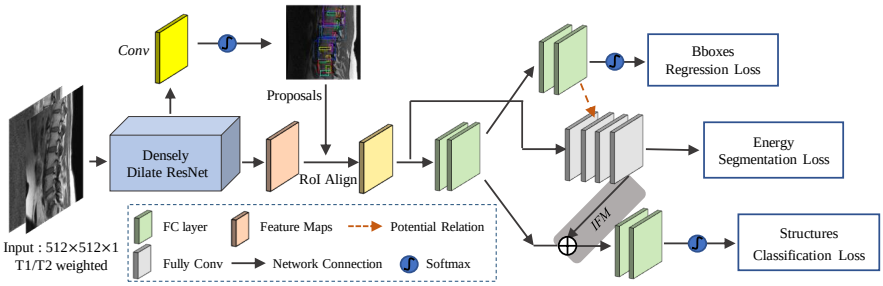


Figure 2: MMCL-Net Architecture. MMCL-Net modules not only cooperate through multi-tasks through parameter sharing but also globally joint optimize through the connections between modules.

intra-class similarities, as illustrated in Fig. 1(d). (3) Latent correlation: the correlation between multi-task and multiple structures are harder to learn than the simple analysis of one type of structure. From the perspective of multiple learning tasks, multi-task have mutually reinforcing relationships, but there can also be a poor performance task that inhibits other tasks. Moreover, parameter sharing among multiple tasks and hyperparameter settings (learning rate, etc.) of model are also barriers to confirm high-performance model structure (Ren and Jae Lee, 2018).

1.1. Related works

Most of the previous work devoted to a single task, which can be divided into three categories: (1) Detection: several studies utilized traditional image processing methods such as the histogram of oriented gradients (Ghosh et al., 2012), probabilistic models (Corso et al., 2008), and deformable hierarchical model (Zhan et al., 2012) to achieved the localization of vertebra or discs. In addition, the deep convolutional network greatly improves the detection accuracy of the spine structure by acquiring more potential information and advanced features of images(Han et al., 2018b; Liao et al., 2018). (2) Segmentation: several studies adopted boundary regression model (He et al., 2018), topology refinement (Klin-

der et al., 2008), GrowCut (Egger et al., 2017) and adaptive spectral (He et al., 2017b) to accomplished the segmentation of one or two types of spine structure. (3) Classification: Existing studies often designed several manual features (Panjabi et al., 2006) or used feature learning algorithms (Ghosh et al., 2012) to extracted low-level information and implement traditional machine learning classifiers on one spine structure. There are also some studies that used distance metric learning (He et al., 2018; Zhang et al., 2017) to maximize interclass margins. Recently, few new studies on multi-structure and multi-task of the spine. Jamaludin et al. (2017) designed a multi-task VGG architecture developed for the multiple analysis of intervertebral discs. Lu et al. (2018) have achieved high performance for automated vertebral segmentation and spinal stenosis grading used U-Net in spine MRI. Multi-task learning of the spine structure has also been involved in our previous work (Han et al., 2018a). But it only achieved the abnormal detection of the three structures, which is completely different from our idea of using the joint optimization method to improve the performance of the model.

Multi-task learning (Caruana, 1997) provided an effective way to enhance the contextual awareness of the algorithm in a global context, which makes the network more emphasized on auxiliary information. The tasks for natural images, Mask R-CNN (He et al., 2017a) is a classical and effective multi-task method that can be used for detection, segmentation, and classification simultaneously. In contrast to most recent systems(Dai et al., 2016), (Li et al., 2017), (Pinheiro et al., 2015), where mutual promotion between multi-tasks, classification depends on object detections and mask predictions. For medical image processing, multi-task learning can exert the above characteristics better with more complex potential information. Pisov et al. (2018) extended the classification-based approach using multi-task learning with heterogeneous labels and implemented corresponding retrieval system based on tumor images . Asgari et al. (2019) introduced a novel multi-task approach with multi-decoder architecture for multiclass segmentation in retinal OCT images. Bai et al. (2019) proposed a novel way for training a cardiac MR images segmentation network, in which fea-

tures are learned in a self-supervised manner by predicting anatomical positions and do not require extra manual annotation. The above multi-task learning methods have achieved better performance than single tasks. Further, considering the particularity of medical image features representation, we propose that specific features (such as anatomical features) of individual task should be emphasized for multi-task learning methods, and task branches ought not to be isolated completely.

1.2. Proposed framework

We propose the Multi-task Multi-structure Correlation Learning Network (MMCL-Net) that handles the instance segmentation, radiology classification, and targets detection to achieve automated fully analysis of common spine diseases. The proposed method has two main advantages. First, compared with the existing spine image processing methods, MMCL-Net adopts the method of learning by aggregate three tasks in a network, it will facilitate to mining potential information between the various structures of spine. Second, compared with the existing multi-task learning methods, MMCL-Net is not isolated into three tasks/branches strictly. It leverages information fusion among tasks to enhance further the effects of multi-task learning. The structure diagram of MMCL-Net is shown in Fig. 2. MMCL-Net implements dynamic joint optimization between multi-structure multi-task through three encoders corresponding to specific tasks. There are three idiographic modules in the model are applied for local optimization: (1) the Densely Dilated ResNet(DDRN) module explores the correlation among multi-task and multi-scale densely features among multi-structure by densely paralleling and cascading of several atrous convolutional layers with different dilation rates; (2) the Deep Convolution Level Set(DCLS) module integrates level-set into softmax function with energy minimization to partition blurry edges, extract instance region and edge features; (3) in addition to the potential logical relationship between the tasks of detection and segmentation, the Instance Feature Merge(IFM) module completes the second progression between the tasks of segmentation and classification by merging hi-

erarchical global features and semantic local features. Additionally, there are three submodules that play important roles in the MMCL-Net model. Feature Pyramid Network (FPN) (Lin et al., 2017) is designed to fuse high-level semantic information from different layers. The Region Proposal Network (RPN) (Ren et al., 2015) is used to generate the archor to fulfill the Regional Proposal. ROI Align uses bilinear interpolation to obtain image values at pixels with floats, which reducing the misalignment of the detection and segmentation matches through transforms the entire feature aggregation process into continuous operation (He et al., 2017a).

1.3. Contributions

The contributions of our work for the construction of this combined model including:

- A novel methodology that using the joint advantages of progressive multi-task learning and multi-structure learning for stimulating model performance to address the problem of spinal MRI diseases detection and the grading of LNFS is proposed.
- It is the first time that simultaneous automated detection, classification, and segmentation on intervertebral disc, vertebrae, and neural foramen are achieved with higher performance than previous individual tasks.
- The variational level-set method with local energy minimization is integrated into softmax function to promote the perception and discrimination ability for fuzzy edges of the model.

2. Methodology

We propose an efficient and effective architecture called MMCL-Net, which shares a common encoder over three tasks and has three branches that have the progressive logical relationship between them. Each of branches implements a decoder for a given task, as illustrated in Fig. 2. MMCL-Net has three novel

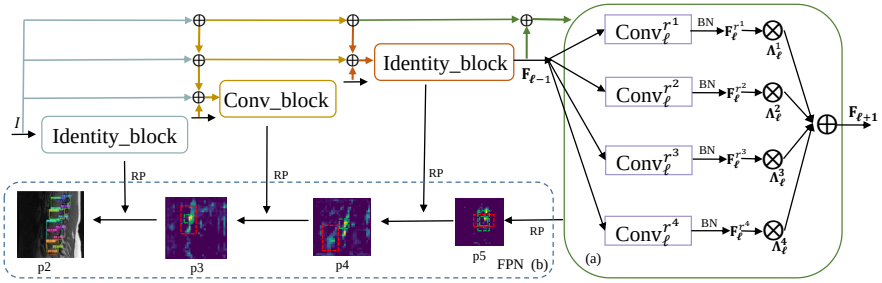


Figure 3: The DDRN module contains layers in three ResNet modules and four parallel dilated convolution layers for weighted summation of activations. The different dilated rates are of 3, 6, 12, and 24, respectively. Addition, the top-down structure feature pyramid network is for proposing anchors.

modules (see Sec. 2.1), advanced optimization algorithm (see Sec. 2.2), and rigorous learning strategy (see Sec. 2.3).

2.1. Three advanced modules

There are three advanced modules in MMCL-Net: (1) The Densely Dilated ResNet module is designed to combat the challenge of extracting significant radiological features from intervertebral discs, vertebrae, and neural foramen, which consist of dense multi-level aggregation dilated residual units. They create a spatially dynamic connection between three structures and three tasks as a committed step. (2) The Deep Convolution Level Set module retains the characteristics of the variational level-set and deep convolution makes MMCL-Net feasible accurately perceive structural blurring edges. (3) The Instance Feature Merge module combines the global features extracted by DDRN and the local features obtained by segmentation. The module improves the ability of MMCL-Net to distinguish between fine-grained instances. The details of three advanced modules are as follows.

2.1.1. Densely dilated ResNet module

The parameter sharing part of the complex model of multi-task is the top priority of the model. The DDRN module as the encoder of the whole model. For one thing, it can simultaneously adaptively optimize model through shared representation learning between multi-task. For another, it extracts features that are densely covered to receptive domains of different sizes. Therefore, two strategies are motivated by DenseNet (Huang et al., 2017): densely paralleling and cascading of several atrous convolutional layers with different dilation rates, as shown in Fig. 3. In dense parallel mode, multiple ResNet layers accept the same input and their outputs are connected, so the resulting output is a sample of inputs with different receptive domain dimensions. In cascade mode, the cascade layers accept the output of the parallel layers, so it can effectively generate larger receptive domains.

According to the basic mathematical model of DenseNet, the ℓ layer receives the feature maps of all preceding layers, $\mathbf{F}_0, ..., \mathbf{F}_{\ell-1}$, as input:

$$\mathbf{F}_\ell = H_\ell([\mathbf{F}_0, \mathbf{F}_1, ..., \mathbf{F}_{\ell-1}]), \tag{1}$$

where $[\mathbf{F}_0, ..., \mathbf{F}_{\ell-1}]$ refers to the concatenation of the feature maps produced in layers $0, ..., \ell-1$. H_ℓ can be a composite function of operations such as Batch Normalization(BN), Rectified Linear Units (ReLU), pooling, or convolution (Conv). Unambiguously classifying different fine-grained structures in an image requires combinations of local and global information. The above features are fused by multi-level of parallel concatenation. Given activations of the previous layer $\mathbf{F}_{\ell-1}$, MMCL-Net capture multi-scale information by processing it in parallel via K convolutional layers with different dilation rates r^k. The K tensors $\mathbf{F}_{\ell-1}^{r^k}$ is shown in Fig. 3(a), each set has the same number of channels M. They detection patterns are at K different scales, then normalizes the K by an element-wise softmax σ for each voxel to add up to one. Let this normalized output be $\mathbf{\Lambda}_\ell = [\mathbf{\Lambda}_l^1, \mathbf{\Lambda}_\ell^2, ..., \mathbf{\Lambda}_\ell^K] \in R^{W\times H\times M}$. Then it is merged in a data-driven manner by introducing a soft attention mechanism (Bahdanau et al., 2014). Thus, the final output of the parallel dilated convolutions layer is

computed by fusing as follows:

$$\mathbf{F}_{\ell+1} = \sum_{k=1}^{K} \mathbf{\Lambda}_{\ell}^{k} \cdot \mathbf{F}_{\ell}^{r^k}, \tag{2}$$

where $\cdot$ denote an element-wise multiplication. The attention weights $\mathbf{\Lambda}_{\ell}^{k}$ are shared across all channels of tensor $\mathbf{F}_{\ell}^{r^k}$ for scale k. The sizes of spinal anatomical structures are varied, while the overall appearance is rather similar, obviously. This shows that scale invariance can be used to standardize model learning capabilities as long as it is used properly. In order to make each filter seeks for structures that similar but of different sizes, the method of sharing parameters between parallel network filters (Yang et al., 2018) so that K does not contain trainable parameters.

2.1.2. Deep convolution level set module

The edges of some spinal structures are weak and blurred, especially the lesion areas. To improve the local performance of the segmentation network, the level-set is combined with the convolutional network framework to extract the deep region and edge features by referring to the ideas of (Duan et al., 2018). This module not only improves MMCL-Net perception of fuzzy edges but contributes to intra-class fine-grained distinguish.

Region features. The segmented spine patch is defined as $u : \Omega \to R^n$, where n is the channel of images. C indicates the edges of the target regions, which divides the given single spine structural patch into two sub-regions $\{\Omega_i\}_{i=1}^{n}$, the foreground (structure area) and the background. The patch $u' = \sum_{i=1}^{n} c_i \psi_i$ is utilized to approximate the original image patch and its label during the segmentation process. Therefore, the summation of the energy of the patch for all pixels is

$$\varepsilon_x(c_i, \varphi) = \int \sum_{i=1}^{n} \lambda_i \int_{\Omega_i} K_\sigma(x-y) \mid I(y) - c_i(x) \mid^2 \psi_i(\varphi(y)) \, \mathrm{d}x\mathrm{d}y, \tag{3}$$

where $x \in \{\Omega_i, i = 1\}$, ψ_i represents the LS feature function as follow

$$\psi_i = \frac{1}{\alpha_i} \prod_{j=1, j\neq i}^{n} (\varphi - j) \quad and \quad \alpha_i = k \prod_{j=1, k\neq i}^{n} (i - k), \tag{4}$$

where K_σ is *Gaussian* kernel function, it can be described as

$$K_\sigma(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{\frac{-|x|^2}{2\sigma^2}}, \tag{5}$$

where $\sigma > 0$ has the characteristic that $K_\sigma(x - y)$ gradually becomes zero in the process of y gradually moving away from x. If the pixel is included by the contour C, the corresponding pixel position has a stronger reflection and returns a smaller value. The level-set region energy is combined with the convolutional neural network and formulate the learning problem as follows: the input training data set is denoted by $S = \{(U_p, R_p, E_p), p = 1, \ldots, N\}$,where samples $U_p = \{u_i^p, j = 1, \ldots, \mid U_p \mid\}$ represent the raw input images, $R_p = \{r_i^p, j = 1, \ldots, \mid R_p \mid\}, r_i^p \in \{1, \ldots, n\}$ and $E_p = \{e_i^p, j = 1, \ldots, \mid E_p \mid\}, e_i^p \in \{1, 1\}$ represent the ground truth regions and binary edge maps for U_p respectively. $\mathbf{W}$ is used to represents all the parameters of the network layer, then the loss function of the region feature extraction is defined as

$$L_R(\mathbf{W}) = -\sum_j \log P_{so}(r_j) \cdot \varepsilon_x(c_i, \varphi), \tag{6}$$

where j is the index of the pixel. The softmax function is utilized to derive the probability of the image U at pixel j from the network.

Edge features. In order to obtain the optimal edge and make the curve as smooth as possible in the evolution of the level-set function. The non-convex regular term is utilized to constrain the edge information in the energy generalization function, and obtains the circumference feature of spine structures to assisted class grading. Then the curve length penalty term is

$$L(\varphi) = \int_C \eta(\mid \bigtriangledown I(x) \mid) \gamma(\mid \bigtriangledown H(\varphi(x) \mid) \, \mathrm{d}x, \tag{7}$$

where $\eta(s) = \frac{1}{1+s^2}$ is the boundary stops the function. The purpose of the non-convex function $\gamma(s) = \frac{s^2}{1+s^2}$ is to reduce the degree of blurring of the *Heaviside*

function at the target contour. Since the *Heaviside* function has jumps at the target contour, the regularization function $H_\epsilon(x)$ of the *Heaviside* function is required for smoothness processing. After combining the above edge feature problem with the network, the loss when extracting edge features is

$$L_E(\mathbf{W}) = -\beta \sum_{j \in Y_+} \log P_{si} \cdot L_{e_j=1}(\phi) - (1-\beta) \sum_{j \in Y_-} \log P_{si} \cdot L_{e_j=0}(\phi), \quad (8)$$

where β is a class-balancing weight. $\beta = |Y_-|/|Y|$ and $1-\beta = |Y_+|/|Y|$, where $|Y_-|$ and $|Y_+|$ are represented as edge and non-edge ground truth label pixels, respectively. P_{si} represents the probability of the boundary position pixel obtained by the network using the sigmoid activation method.

Then, the segmentation problem of the spine structures is minimized by the total energy generalization function of the entire segmented region containing the edge profile transformed by the level-set

$$\min_{c_i, \varphi} E = \varepsilon_x(c_i, \varphi) + L(\varphi). \quad (9)$$

After the network is trained, the validation set and obtain the joint region and edge probability maps from the last convolutional layer. The overall loss function for the segment network branch is

$$\mathbf{W}^* = \arg\min(L_R \mathbf{W} + \alpha L_E(\mathbf{W})) \quad (10)$$

where $L_R(\mathbf{W})$ and $L_E(\mathbf{W})$ are respectively the region associated cross-entropy loss that enables the network to learn region and edge features. Weight α is denoted to balance the two categories. The application of DCLS allows the model to learn more anatomical features of the spine structures. Then, the segmented morphological features are fused into the global features for intra-class grading of the structure.

2.1.3. Instance feature merge module

The segmentation area and the perimeter of the boundary are important characteristic parameters for accurate identification of the structure area, especially for medical fine-grained images with very small differences within the class.

Therefore, IFM directly merges the last convolutional layer segmented by the instance into the global features extracted by the DDRN. As shown in Fig. 4, the joint region and edge probability maps are randomly cropped and are then down-sampled to the same size as the DDRN feature maps, and then a 3x3 convolution is applied before connecting to the output of the merged layer. Subsequently, a 3x3 convolution is applied to the concatenated block, and then the output is used for the residual summation operation instead of the input tensor in the conventional method.

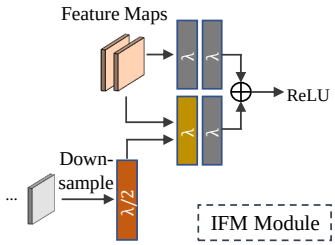


Figure 4: The IFM module structure. $\lambda = 256$ refer to the output depth.

$$\mathbf{F}_{all} = \mathbf{F}_{l+1}(x; \mathbf{W}) + \mathbf{F}_{downpooling}(x; \mathbf{W}^*) \tag{11}$$

2.2. Optimization

The minimization process of Eq.(9) needs to differentiate the variables. We use Nesterov algorithm to optimize the energy generalization function, and minimize Eq.(9) to get the evolution equation about the level-set function φ as

$$\begin{aligned} \frac{\partial E}{\partial \varphi} = & \sum_{i=1}^{n} \lambda_i \int_{\Omega_i} K_\sigma(x-y) \mid I(y) - c_i(x) \mid^2 \frac{\partial \psi_i}{\partial \varphi} \mathrm{d}y \\ & - \eta \mid \bigtriangledown I \mid \bigtriangledown \mid \frac{\bigtriangledown H_\varepsilon(\varphi)}{(1+ \mid \bigtriangledown H_\varepsilon(\phi) \mid^2)^2} \mid \end{aligned} \tag{12}$$

At the steady state of Eq.(12), the local or global minimum of Eq.(9) can be found. But the energy formula Eq.(9) is non-convex, the crossover iteration method is used to make the model converge to the solution we want.

MMCL-Net uses multi-task regression learning on both location and classification or grading tasks. Therefore, the multi-task regression loss as follows

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{W_d}) = \sum_{m=1}^{M} \sum_{i \in \chi} \alpha_m l^m(X_i, Y_i | \boldsymbol{W}) + \sum_{i \in S^{M+1}} l^{M+1}(X_i, Y_i | \boldsymbol{W}, \boldsymbol{W_d}), \tag{13}$$

where l^{M+1} and S^{M+1} are the loss and training samples for the detection branch. l^{M+1} represents the summation of cross-entropy loss for classification and the smooth L_1 loss for proposals regression. The detection branch shares some of the encoder network parameters $\boldsymbol{W}$ and adds $\boldsymbol{W_d}$ parameters.Common optimization strategy is used for these parameters, i.e. $(\boldsymbol{W}^*, \boldsymbol{W}_d^*) = \arg\min \mathcal{L}(\boldsymbol{W}, \boldsymbol{W_d})$. Stochastic gradient descent method is utilized to optimize the mixed loss function, the whole optimization process is fully automated.

2.3. Implementation details

We used standard five-fold cross-validation for performance evaluation and comparison due to the limited volume of data sets. This method enables all samples to be fully utilized and each sample used only once for testing. Our framework is implemented based on the open-source software library TensorFlow version 1.9.0. The momentum and weight decay as 0.9 and 0.0001 respectively. Learning rate is 0.0001 and optimizer is SGD. Training batch size is 2 and maximum iteration is 300 using a Nvidia Quadro K2200 GPU with cuDNN v7.0 and Intel I7 CPU.

3. Experiments

3.1. Evaluation metric and data

We conducted the performance test of MMCL-Net using the sample dataset contains 200 clinical patients(Avg 60 yrs), and selected the middle sagittal slice in a T1/T2-weighted sequence MR image of each patient to compose a data set. Specifically, this dataset has about 1200 vertebrae (518 normal, 682 abnormal), 1200 discs (627 normal, 573 abnormal), and 1000 neural foramina (normal 622, slight 211, marked 117, severe 50) respectively. The intervertebral discs and vertebra were labeled to normal and abnormal according to clinical criteria for T1/T2-weighted 2D image data of 200 clinical patients, denoted by D0, D1, V0, and V1 respectively, as shown in Fig. 5. According to the Wildermuth qualitative grading system, the neural foramina are divided into four grades: normal,

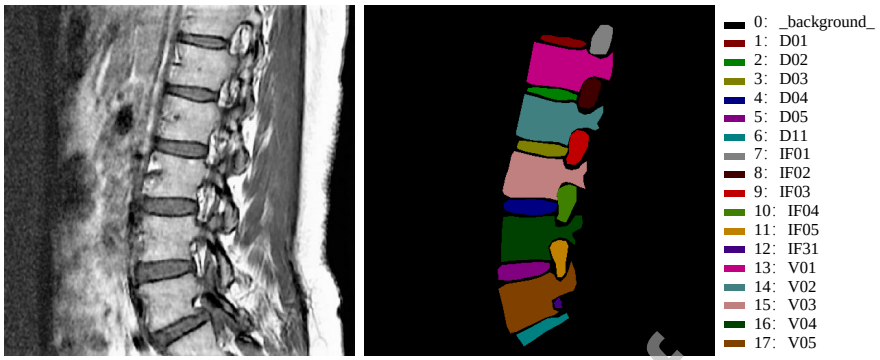


Figure 5: The labels of spinal image segmentation and classification. D-, V-, and IF- represent intervertebral disc, vertebral, and lumbar neural foramen respectively.

sight, marked and severe, which are represented by IF0, IF1, IF2, and IF3 respectively. The ground truths were annotated by one experienced physician. When training, random cropping and randomly selecting of different subset of instances for each epoch are used to augment the dataset.

The segmentation performance was verified in terms of pixel intersection-over-union (IoU) averaged across every class. The accuracy and harmonic mean are utilized for performance evaluation. Three indicators including accuracy, sensitivity and harmonic mean are used in the detection.

3.2. Results

The effectiveness and superiority of MMCL-Net in the analysis of multiple spine structures have been demonstrated by extensive experimental results. As illustrated in Fig. 6 and Fig. 7, we visualized the classification labels and segmentation labels of the spine structure, the detection results of Mask R-CNN, DDRN+DCLS(without IFM), and MMCL-Net from left to right. MMCL-Net not only detect exactly and distinguish abnormal structures but also discover their parthenogenesis correlation between disc, vertebra and neural foramina (the instances showing by red boxes in the left column). Several indicators are selected from two aspects to evaluate the performance of the proposed model

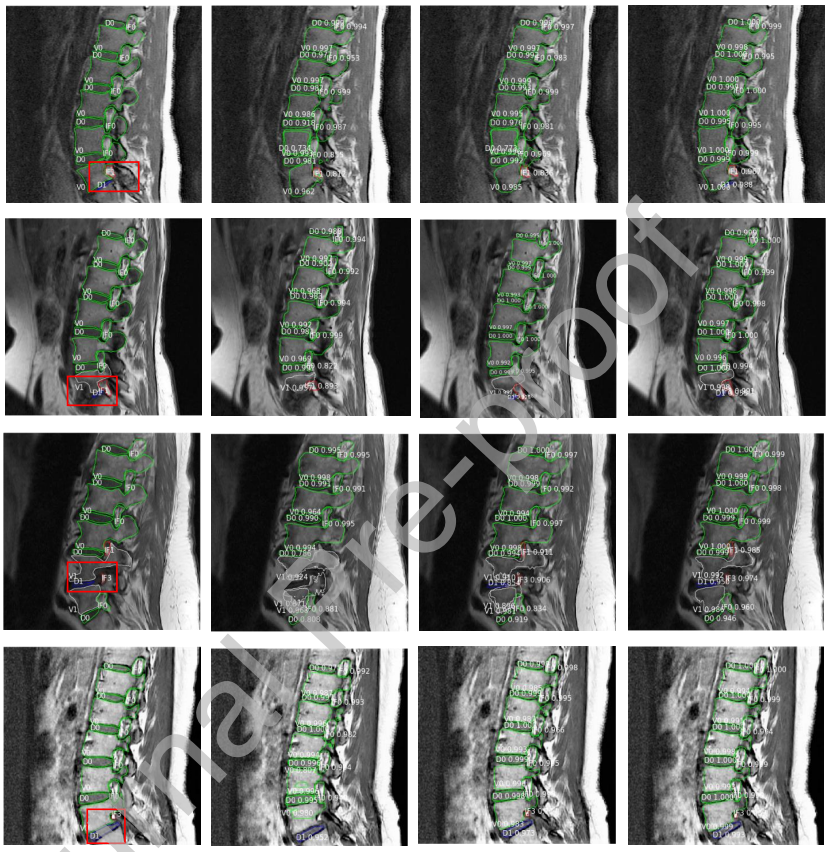


Figure 6: Global visualization of segmentation and classification results, where the normal structure as green and the abnormal structure as other colors for. D0, V0, and IF0 represent normal status of the three structures (in green), D1 and V1 represent abnormal intervertebral disc (in blue) and abnormal vertebral (in gray), respectively. IF1, IF2, and IF3 represent lumbar neural foramen stenosis slight, marked, and severe respectively (in red).

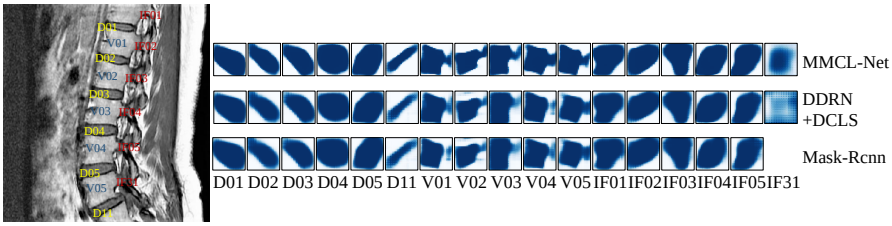


Figure 7: Visualization of spinal structure segmentation masks by MMCL-Net with(the top row) and without(the middle row) DCLS. The last line indicates the segmentation result of Mask R-CNN. The mask size is 28×28. The vacancy at the position of the coordinate (IF31, Mask R-CNN) means that Mask R-CNN does not detect this position.

and compare it with the latest algorithms and state-of-the-art researches.

We conducted statistical analysis to ensure that the results are interpreted correctly and whether the apparent relationship in the data reflects a true relationship in the population. Firstly, the paired t-test between Mask R-CNN and MMCL-Net at a 5% significance level with *p-value* of 0.014 for three tasks, which clearly indicates the improvements of our method are statistically significant. Then, the *p-values* among FCN and MMCL-Net for segmentation task, Faster RCNN and MMCL-Net for classification and detection are also less-than 0.05 proving the achieved significance. The specific statistical results are shown in Table 1 and Table 2.

Table 1: Dice overlap metric between the automated and ground truth for evaluating segmentation performance. Best results are highlighted with bold.

Method	Disc		Vert.		NF			
	Normal	*Abnormal*	*Normal*	*Abnormal*	*Normal*	*Slight*	*Marked*	*Severe*
Spine-GAN (Han et al., 2018b)	0.873	0.840	0.930	0.810	-	-	-	-
FCN (Long et al., 2015)	0.8142	0.7256	0.8701	0.7910	0.8262	0.7644	0.7551	0.7034
Mask R-CNN (He et al., 2017a)	0.8671	0.7878	0.9153	0.8361	0.8637	0.8191	0.7824	0.7122
DDRN (without FPN)	0.8395	0.7644	0.8943	0.8278	0.8469	0.8033	0.7937	0.7606
DDRN (without FPN)+DCLS	0.8587	0.8156	0.9091	0.8489	0.8574	0.8111	0.8105	0.7878
MMCL-Net	**0.9231**	**0.8962**	**0.9474**	**0.9178**	**0.9128**	**0.8993**	**0.8789**	**0.8222**

3.2.1. Segmentation performance

The segmentation performance was verified by calculating the Dice overlap metric, as shown in Table 1. In order to verify the effectiveness of each module, we replaced the network structure in FCN with DDRN module (without FPN) for the segmentation task. Although the *DDRN (without FPN)* structure does not have the detection function, its segmentation performance is slightly higher than FCN. As shown in the results of the *DDRN (without FPN)* column in Table 1, the improvement of segmentation performance range in different spine structures is between 0.0207 and 0.0574. The highest improvement in the segmentation result is the Severe grade of NF class, due to the DDRN structure has a better ability to capture fine structures than other ordinary convolutional networks. Further, FCN was compared with the classic multi-task network Mask R-CNN, the results in Table 1 demonstrate that multi-task structure is more precisely than single-task for segment multiple structures of spine. For the function-sharing network structure, the average improvement is 0.0417. Unlike networks that are used solely for segmentation, Mask R-CNN is actively trained and benefits from the better structure. Simultaneously, the segmentation performance of Mask R-CNN for abnormal structures is much lower than MMCL-Net. The improvement of segmentation results is mainly lesion structures of spine. The analysis demonstrates that MMCL-Net has a lower classification error rate than Mask R-CNN, and the improvement of classification greatly promotes the improvement of segmentation performance.

The experimental results of MMCL-Net are also compared with spine-GAN proposed in (Han et al., 2018b), which code is publicly available. The predominance of MMCL-Net exceed this kind of generated antagonism network is that correlation framework implements instance segmentation of the target is based on the accurate detection of the IoU region, instead of segment the entire image directly. The benefit of this is to reduce intra-class interference between different structures so that the model only considers different grades between the same structure. In summary, MMCL-Net achieves the best segmentation performance

on all types of target spine structures. As shown in Fig. 6, MMCL-Net is able to accurately segment structures that have severe deformation degradation. The segmentation result is visualized as shown in Fig. 7. On the one hand, DCLS enables MMCL-Net to effectively identify tiny lesion structures and segment them accurately. On the other hand, such a good segmentation result is also due to the accurate detection of the MMCL-Net detection branch. The design of the global network is richly endowed by nature for the task of complex medical structures.

From the perspective of mistake, the segmentation fault of MMCL-Net is mainly registered as the areas where the edges are not obvious and where the locations are wrong. Due to the MR image has only a single channel and the texture features of the spine structures are not obvious. Use 3D MR image would certainly improve the situation, and more samples are required as well.

Table 2: The comparisons between MMCL-Net and state-of-the-art methods in the classification and detection tasks. Best results are highlighted in bold.

Method	Classification				Detection				mAP
	Acc.			F_1	Acc.			Sen.	
	Disc	Vert.	NF		Disc	Vert.	NF		
TASRL (He et al., 2018)	-	-	**0.943**	**0.898**	-	-	**0.9992**	**1.00**	-
DMML-Net (Han et al., 2018a)	0.814	0.789	-	-	0.827	0.835	0.806	-	0.811
DEEP SPINE (Duan et al., 2018)	-	-	0.781	0.8047	-	-	-	-	-
Faster RCNN (Ren et al., 2015)	0.7924	0.8251	0.7444	0.7556	0.8968	0.9134	0.8742	0.8860	0.8274
Mask R-CNN (He et al., 2017a)	0.8235	0.8443	0.7819	0.8022	0.9241	0.9434	0.9028	0.9142	0.8334
Faster RCNN (with DDRN)	0.8545	0.8762	0.8441	0.8362	0.9861	0.9814	0.9824	0.9837	0.8601
-DCLS*	0.8558	0.8742	0.8426	0.8323	0.9859	0.9800	0.9818	0.9844	0.8616
MMCL-Net	**0.9113**	**0.9312**	0.9187	0.8975	**0.9890**	**0.9810**	0.9827	0.9869	**0.9187**

* *represents mask branch in the Faster RCNN (with DDRN) model replaced by DCLS module.*

3.2.2. Detection and classification performance

Table 2. shows the results of MMCL-Net, the best algorithms available from other research, and the models that before MMCL-Net adding each optimization module. To validate the impact of individual module on detection and classification tasks, we compared the performance of different models include

DMML-Net (one stage), Faster RCNN (two stages), and Mask R-CNN (multi-task) for detection and classification. The multi-task model has much better recognition and classification performance than the single task models. In addition, the performance of Faster RCNN(with DDRN) is better than Faster RCNN as shown in Table 2, which proves the advantage and potentiality of DDRN module in identifying small-scale spinal structures. Furthermore, the results of the three models Faster RCNN(with DDRN), DDRN+DCLS(without IFM) and MMCL-Net demonstrate that DCLS only contributes to the segmentation task of MMCL-Net and has little impact on classification and detection. However, the performance of MMCL-Net on the classification task is greatly improved after the IFM model has added, which illuminates that IFM module can assist other tasks in effectively utilizing the anatomical features from the segmentation branch.

The identification of abnormal structures can reflect better the discriminative performance of MMCL-Net on spinal diseases. From the average estimation shown in Table 2. and visualization shown in Fig. 6, can observe following: (1) MMCL-Net outperforms conventional models in detection and classification, which can be ascribed to the relevance of DDRN preliminary estimation that modeling the dependencies of multi-structures and multi-tasks; (2) the accurate detection and recognition of the deformation structure also depends on the merging of various local and global features. The comparison of the experimental results in Table 2 and Fig. 8 shows that MMCL-Net has outstanding predominance in detection of abnormal intervertebral discs and vertebra, but the performance in the grading of neural foramen is slightly lower than that of TASRL (He et al., 2018) in all indexes. TASRL is utilized saliency-biased Ncuts for efficient localization and proposed to learn the intra-label robust representation for tolerating the intra-object difference in localization and intra-grade difference. And TASRL has obvious advantages for small sample sets comparing to convolutional neural networks.

From the perspective of model detection and classification mistake, classification errors assemble mainly in the identification of abnormal spinal struc-

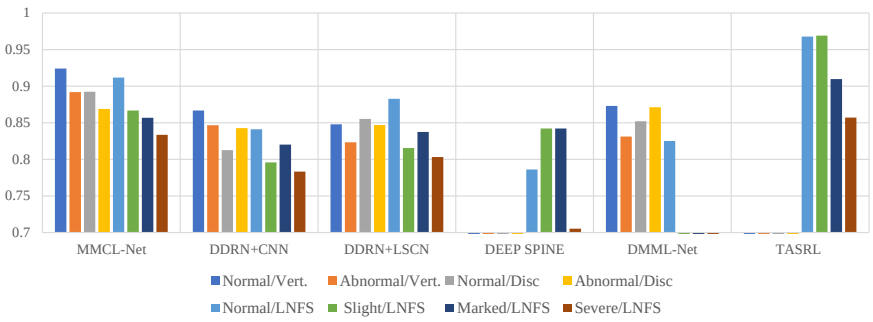


Figure 8: The bar diagram of recognition capabilities from different structure models in abnormal vertebra, abnormal disc, and LNFS.

tures. The intra-class differences between abnormal and normal structures is very small, which is one of the difficulties in the identification of pathological structures in medical images. Detection errors are similar to segmentation tasks.

4. Conclusion and discussion

The innovative multi-structure and progressive multi-task joint optimization composite network MMCL-Net can simultaneously and steadily implement automatic detection, segmentation, and classification of multiple spinal diseases. The local optimization of MMCL-Net is achieved through a densely extended ResNet module and a DCLS module that for optimizing segmentation performance, and an IFM module that fuses the extracted anatomical features of the level-set into deep learning features. Through the verification of MR images of 200 clinical patients, MMCL-Net can not only detect accurately, segment and identify the pathological correlation of multiple spinal structures, but also show superior performance in the grading problem of lumbar neural foramen stenosis. The new automated composite model is not just helpful for clinical pathogenesis-based diagnosis of spinal diseases, but brings new directions for collaborative multitasking due to the similar concepts can be applied to other applications.

sectionInformation Sharing StatementSource data can be directly downloaded at https://drive.google.com/file/d/1W64dSL-DOCUt3JOEF0csnqddmMHF4-7/view?usp=sharing.

Credit Author Contribution

Yanfei Hong: Conceptualization, Methodology, Software, Writing - Original Draft Benzheng Wei: Supervision, Project administration, Resources. Zhongyi Han: Validation, Formal analysis, Writing - Review & Editing Xiang Li: Visualization, Data Curation Yuanjie Zheng: Supervision, Validation Shuo Li: Investigation

Declaration of Competing Interest

The authors declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partly funded by Natural Science Foundation of China (No.61872225); the Natural Science Foundation of Shandong Province (No.ZR2015FM010); the Project of Science and technology plan of Shandong higher education institutions Program (No.J15LN20); the Project of Shandong Province Medical and Health Technology Development Program (No.2016WS0577).

References

References

Asgari, R., Orlando, J.I., Waldstein, S., Schlanitz, F., Baratsits, M., Schmidt-Erfurth, U., Bogunović, H., 2019. Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography. arXiv preprint arXiv:1906.07679 .

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. arXiv preprint arXiv:1907.02757 .

Caruana, R., 1997. Multitask learning. Machine learning 28, 41–75.

Corso, J.J., RajaS, A., Chaudhary, V., 2008. Lumbar disc localization and labeling with a probabilistic model on both pixel and object features, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 202–210.

Dai, J., He, K., Sun, J., 2016. Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158.

Duan, J., Schlemper, J., Bai, W., Dawes, T.J., Bello, G., Doumou, G., De Marvao, A., ORegan, D.P., Rueckert, D., 2018. Deep nested level sets: Fully automated segmentation of cardiac mr images in patients with pulmonary hypertension, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 595–603.

Egger, J., Nimsky, C., Chen, X., 2017. Vertebral body segmentation with growcut: Initial experience, workflow and practical application. SAGE open medicine 5, 2050312117740984.

Ghosh, S., Malgireddy, M.R., Chaudhary, V., Dhillon, G., 2012. A new approach to automatic disc localization in clinical lumbar mri: combining machine learning with heuristics, in: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on, IEEE. pp. 114–117.

Han, Z., Wei, B., Leung, S., Nachum, I.B., Laidley, D., Li, S., 2018a. Automated pathogenesis-based diagnosis of lumbar neural foraminal stenosis via deep multiscale multitask learning. Neuroinformatics , 1–13.

Han, Z., Wei, B., Mercado, A., Leung, S., Li, S., 2018b. Spine-gan: Semantic segmentation of multiple spinal structures. Medical image analysis 50, 23–35.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017a. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

He, X., Leung, S., Warrington, J., Shmuilovich, O., Li, S., 2018. Automated neural foraminal stenosis grading via task-aware structural representation learning. Neurocomputing 287, 185–195.

He, X., Zhang, H., Landis, M., Sharma, M., Warrington, J., Li, S., 2017b. Unsupervised boundary delineation of spinal neural foramina using a multifeature and adaptive spectral segmentation. Medical image analysis 36, 22–40.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks., in: CVPR, p. 3.

Jamaludin, A., Kadir, T., Zisserman, A., 2017. Spinenet: Automated classification and evidence visualization in spinal mris. Medical image analysis 41, 63–73.

Klinder, T., Wolz, R., Lorenz, C., Franz, A., Ostermann, J., 2008. Spine segmentation using articulated shape models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 227–234.

Kokkinos, I., 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory., in: CVPR, p. 8.

Lee, S., Lee, J.W., Yeom, J.S., Kim, K.J., Kim, H.J., Chung, S.K., Kang, H.S., 2010. A practical mri grading system for lumbar foraminal stenosis. American Journal of Roentgenology 194, 1095–1098.

Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y., 2017. Fully convolutional instance-aware semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2359–2367.

Liao, H., Mesfin, A., Luo, J., 2018. Joint vertebrae identification and localization in spinal ct images by combining short-and long-range contextual information. IEEE transactions on medical imaging 37, 1266–1275.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Lu, J.T., Pedemonte, S., Bizzo, B., Doyle, S., Andriole, K.P., Michalski, M.H., Gonzalez, R.G., Pomerantz, S.R., 2018. Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. arXiv preprint arXiv:1807.10215 .

Panjabi, M.M., Maak, T.G., Ivancic, P.C., Ito, S., 2006. Dynamic intervertebral foramen narrowing during simulated rear impact. Spine 31, E128–E134.

Pinheiro, P.O., Collobert, R., Dollár, P., 2015. Learning to segment object candidates, in: Advances in Neural Information Processing Systems, pp. 1990–1998.

Pisov, M., Makarchuk, G., Kostjuchenko, V., Dalechina, A., Golanov, A., Belyaev, M., 2018. Brain tumor image retrieval via multitask learning. arXiv preprint arXiv:1810.09369 .

Rajaee, S.S., Bae, H.W., Kanim, L.E., Delamarter, R.B., 2012. Spinal fusion in the united states: analysis of trends from 1998 to 2008. Spine 37, 67–76.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Ren, Z., Jae Lee, Y., 2018. Cross-domain self-supervised multi-task feature learning using synthetic imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 762–771.

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3684–3692.

Zhan, Y., Maneesh, D., Harder, M., Zhou, X.S., 2012. Robust mr spine detection using hierarchical learning and local articulated model, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 141–148.

Zhang, Q., Bhalerao, A., Hutchinson, C., 2017. Weakly-supervised evidence pinpointing and description, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 210–222.

Biography

Yanfei Hong is a master student of Shandong University of Traditional Chinese Medicine. Her major research interests include machine learning and medical image analysis.

Benzheng Wei received the B.S. degree in computer science from School of Computer Science at Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from School of Computer Science and Technology at Shandong University, Jinan, China, in 2007, and the Ph.D. degree in precision instrument and machinery from College of Automation Engineering at Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. Dr.Wei is a professor with Shandong University of Traditional Chinese Medicine. He is also acting as a director at both the Center for Medical Artificial Intelligence and the Computational Medicine Lab of Shandong University of Traditional Chinese Medicine. His current research interests are in artificial intelligence, medical image analysis and computational medicine. Dr. Wei has published over 80 papers in refereed international leading journals/conferences such as Medical Image Analysis, Neuroinformatics, Neurocomputing and MICCAI.

Zhongyi Han is a PhD student of Shandong University. His major research interests include machine learning and data mining.

Xiang Li received his B.E. degree from School of Information Engineering, Shandong Youth University of Political Science, Jinan, China, in 2018. Currently, he is pursuing his master degree at Shandong University of Traditional Chinese Medicine. His main research interests include machine learning and medical image analysis.

Yuanjie Zheng was a Senior Research Investigator with the Perelman School of Medicine, University of Pennsylvania. He is now a full professor of Shandong Normal University of China and serving as the dean of the School of Information Science and Engineering. His research interests include medical image analysis, translational medicine, computer vision and computational photography.

Shuo Li received his Ph.D. degree in computer science from Concordia University in 2006, Montreal, QC, Canada. He is a Research Scientist and Project Manager at GE Healthcare, London, ON, Canada. He is also an adjunct research professor at the University of Western Ontario and adjunct scientist at the Lawson Health Research Institute. He is currently leading the Digital Imaging Group of London as the Scientific Director. His current research interests are in medial image analysis, with a main focus on automated medial image analysis and visualization.