Medical Image Analysis 55 (2019) 103–115

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Direct automated quantitative measurement of spine by cascade amplifier regression network with manifold regularization

Shumao Pang[a], Zhihai Su[b], Stephanie Leung[c,d], Ilanit Ben Nachum[c,d], Bo Chen[c,d], Qianjin Feng[a,*], Shuo Li[c,d]

[a] Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou, 510515, China
[b] Department of Spinal Surgery, the Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, 519000, China
[c] Department of Medical Imaging, Western University, ON, Canada
[d] Digital Image Group, London, ON, Canada

ARTICLE INFO

Article history:
Received 30 June 2018
Revised 25 February 2019
Accepted 17 April 2019
Available online 22 April 2019

Keywords:
Deep learning
Spine
Local linear representation
Manifold learning

ABSTRACT

Automated quantitative measurement of the spine (i.e., multiple indices estimation of heights, widths, areas, and so on for the vertebral body and disc) plays a significant role in clinical spinal disease diagnoses and assessments, such as osteoporosis, intervertebral disc degeneration, and lumbar disc herniation, yet still an unprecedented challenge due to the variety of spine structure and the high dimensionality of indices to be estimated. In this paper, we propose a novel cascade amplifier regression network (CARN) with manifold regularization including local structure-preserved manifold regularization (LSPMR) and adaptive local shape-constrained manifold regularization (ALSCMR), to achieve accurate direct automated multiple indices estimation. The CARN architecture is composed of a cascade amplifier network (CAN) for expressive feature embedding and a linear regression model for multiple indices estimation. The CAN produces an expressive feature embedding by cascade amplifier units (AUs), which are used for selective feature reuse by stimulating effective feature and suppressing redundant feature during propagating feature map between adjacent layers. During training, the LSPMR is employed to obtain discriminative feature embedding by preserving the local geometric structure of the latent feature space similar to the target output manifold. The ALSCMR is utilized to alleviate overfitting and generate realistic estimation by learning the multiple indices distribution. Experiments on T1-weighted MR images of 215 subjects and T2-weighted MR images of 20 subjects show that the proposed approach achieves impressive performance with mean absolute errors of 1.22 ± 1.04 mm and 1.24 ± 1.07 mm for the 30 lumbar spinal indices estimation of the T1-weighted and T2-weighted spinal MR images respectively. The proposed method has great potential in clinical spinal disease diagnoses and assessments.

1. Introduction

The quantitative measurement of the spine (i.e., multiple indices estimation of heights, widths, areas, and so on for the vertebral body and disc) is a practical means of clinical spinal disease diagnoses and assessments, such as osteoporosis, intervertebral disc degeneration, and lumbar disc herniation. Among these indices to be estimated, the vertebral body height (VBH) and intervertebral disc height (IDH) are the most valuable for these spinal diseases diagnoses and assessments. As shown in Fig. 1, the 30 estimated indices for the lumbar spine include 15 VBHs and 15 IDHs. Each vertebral body (intervertebral disc) contains 3 VBHs (IDHs) including anterior, middle, and posterior VBHs (IDHs). In clinical practice, the VBHs can be used to assess the vertebral fracture risk for the osteoporotic patients (McCloskey et al., 2012; Tatoń et al., 2014) based on the fact that the VBHs are correlated with the bone strength. Furthermore, the IDH decreases with the intervertebral disc degeneration (Jarman et al., 2014; Salamat et al., 2016) and lumbar disc herniation (Tunset et al., 2013).

Automated quantitative measurement of the spine is of significant clinical importance due to several advantages including, time-saving, reproducibility, and higher consistency compared with manual quantitative measurement but remains as an exceedingly intractable task due to the following challenges:

- It is difficult to obtain expressive feature embedding for such complex regression problem due to the high dimensionality of estimated indices (as shown in Fig. 1(a)).

* Corresponding author.
E-mail addresses: 1271992826@qq.com (Q. Feng), slishuo@gmail.com (S. Li).

https://doi.org/10.1016/j.media.2019.04.012
1361-8415/

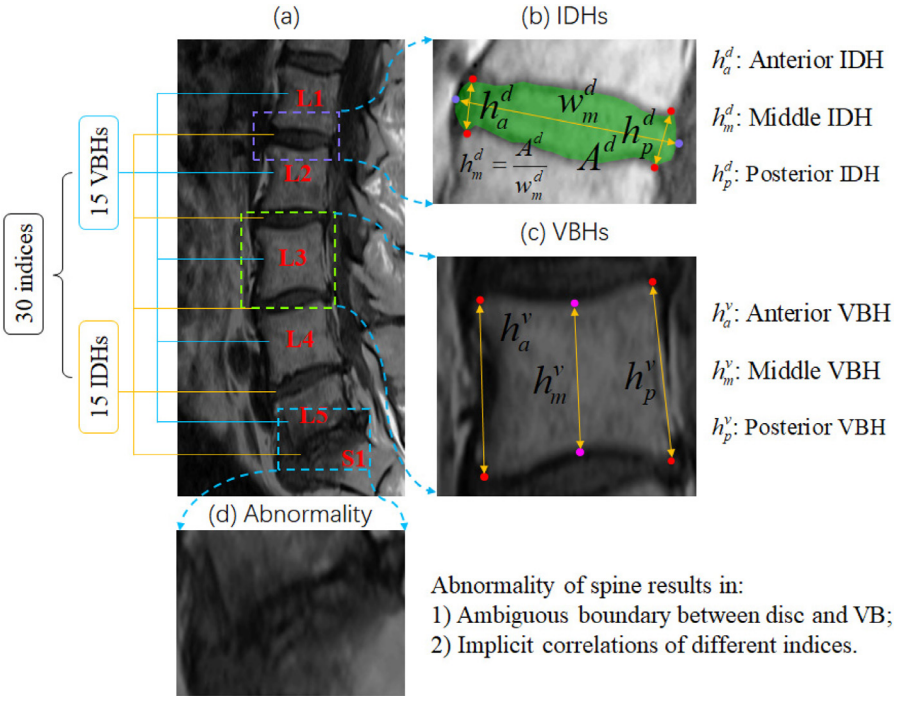


Fig. 1. The challenges of quantitative measurement of the spine detailed in (a) Illustration of 30 indices to be estimated; (b) Three heights for each disc; (c) Three heights for each vertebral body; (d) Ambiguous boundary between disc and VB and implicit correlations of different indices due to spinal abnormality.

- Discriminative feature embedding is intractable to be generated due to the excessive ambiguity of the boundary between vertebral body (VB) and intervertebral disc for abnormal spine (as shown in Fig. 1(d)).
- The implicit correlations between different estimated indices are difficult to be captured (as shown in Fig. 1(d), the heights of the abnormal disc and the heights of adjacent vertebral body are correlated because disc abnormality leads to simultaneous changes of IDH and the adjacent VBH).
- The complex relationship between the spinal images and the estimated indices arises from the variability of images. Images with the same estimated indices often exhibit great variability due to inter-subject variations.
- Insufficient labeled data, which possibly results in overfitting.

1.1. Related work

Existing relevant works for multiple indices estimation of the spine fall into three categories: (1) Manual measurements; (2) automated segmentation; (3) direct estimation.

1.1.1. Manual measurements

Manual measurements aim to quantify the spine by manually measuring the disc height in vitro (Brinckmann and Grootenboer, 1991), detecting the landmark of the spine (Tunset et al., 2013; Videman et al., 2014) from MRI, and segmenting the disc and vertebral body from MRI (Videman et al., 2014). These manual methods are limited in clinical practice because they are time-consuming, tedious, nonreproducible, and susceptible to high inter-observer variability.

1.1.2. Automated segmentation

Automated segmentation-based methods focus on segmenting the intervertebral disc or vertebral body by active shape models (Castro et al., 2012), multi-atlas based models (Wang and Forsberg, 2016), superpixels based models (Barbieri et al., 2015), and deep learning based models (Korez et al., 2017). Although these methods achieve accurate segmentation of the intervertebral disc and vertebral body, the obtained segmentation is incapable of directly computing the required estimated indices.

1.1.3. Direct estimation

In recent years, an increasing number of approaches emerged in the direct quantitative measurement of anatomical structures without the need for segmentation. These methods have achieved great performance in quantitative estimation such as cardiac volume (Xue et al., 2017; Zhen et al., 2016; 2014) and spinal curvature (Wu et al., 2017; Sun et al., 2017). Zhen et al. (2014) used Multi-features and regression forests (Multi-features+RF) to jointly estimate the cardiac bi-ventricular volumes. Zhen et al. (2016) adopted Multi-scale convolutional deep belief network to learn unsupervised cardiac image representation and regression forests (MCDBN+RF) to generate bi-ventricular volumes estimation. Xue et al. (2017) utilized a convolutional neural network (CNN) and recurrent neural network in conjunction with both temporal and spatial information for full quantification of left ventricle. Sun et al. (2017) exploited histogram of oriented gradient descriptor (Dalal and Triggs, 2005) and structured support vector regression (HOG+SSVR) to improve the performance of spinal curvature assessment by exploiting the intrinsic inter-output correlation under the $l2$, 1-norm regularization and preserving the local geometrical structure invariance via manifold regularization.

Although these methods achieved promising performance in the quantification of the cardiac volume and spinal curvature, they are incapable of achieving quantitative measurement of the spine since they suffer from the following limitations. 1) Lack of expressive and discriminative feature representation. The hand-crafted features are not capable of capturing task-aware spinal structures robustly. Traditional CNN (Simonyan and Zisserman, 2014) is incapable of generating an expressive and discriminative feature for multiple indices estimation because CNN possibly loses effective feature due to the lack of an explicit structure for feature reuse. 2) Incapability of learning the estimated indices distribution, which will lead to unreasonable estimation and overfitting.

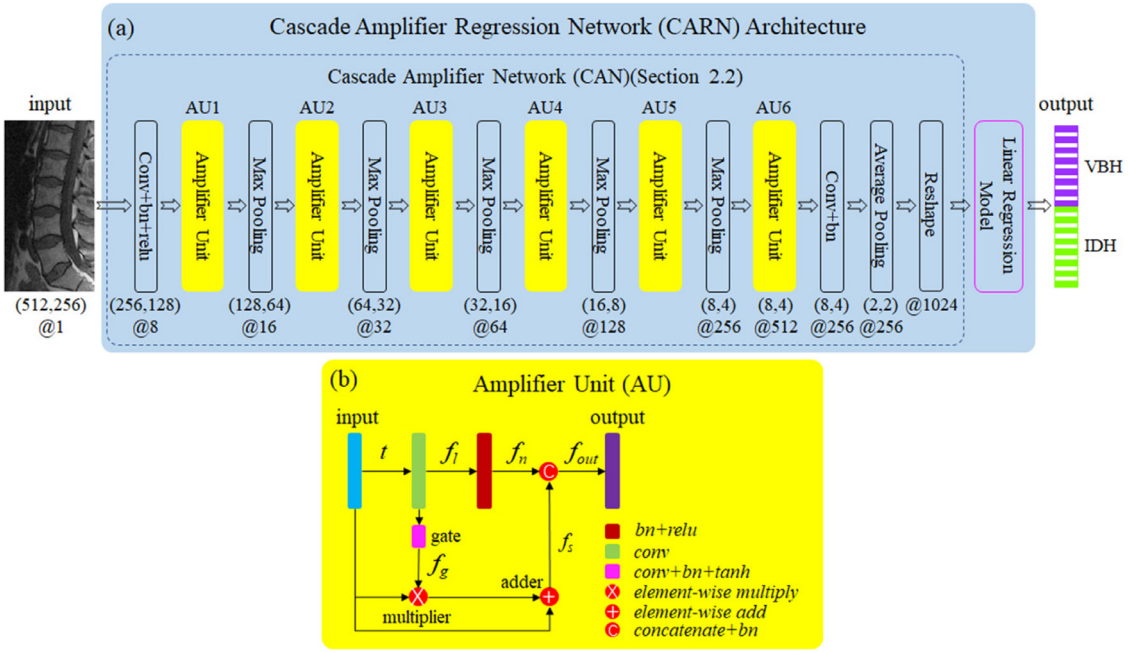


Fig. 2. (a) Overview of CARN architecture, including CAN for expressive feature embedding and a linear regression model for multiple indices estimation. The numbers in brackets at the bottom of (a) denote the feature maps size and @# denotes the number of feature maps. (b) AU for selective feature reuse between adjacent layers.

1.2. Overview of the proposed method

In this study, a cascade amplifier regression network (CARN) with manifold regularization is proposed for quantitative measurement of the spine from MR images. The CARN architecture is comprised of a cascade amplifier network (CAN) for expressive feature embedding and a linear regression model for multiple indices estimation; the manifold regularization including local structure-preserved manifold regularization (LSPMR) and adaptive local shape-constrained manifold regularization (ALSCMR) is proposed to construct the loss function. In the CAN, amplifier unit (AU) aims to reuse the selected feature between adjacent layers. As shown in Fig. 2 (b), the AU generates the selected feature by stimulating the effective feature of the anterior layer but suppressing the redundant feature. The selected feature is reused in the posterior layer by a concatenation operator. CAN reuses multi-level features selectively for representing complex spine, thus an expressive feature embedding is obtained. Using the CAN, the MR images are embedded into a latent feature space. The high dimensional indices lie in a target output manifold due to the correlations between these indices. To take advantage of the relationship between the latent feature space and target output manifold, the LSPMR is proposed to generate a discriminative feature embedding which preserves the local geometrical structure of the target output manifold. Additionally, the ALSCMR is designed to restrict the output of the CARN to the target output manifold. As a result, the distribution of the estimated indices is close to the real distribution, which reduces the impact of outliers and alleviates overfitting. Combining the expressive and discriminative feature embedding produced by CAN and LSPMR with ALSCMR, a simple linear regression model, i.e., a fully connected network, is sufficient to produce accurate estimation results.

1.3. Contributions

The main contributions are as follows:

- A novel regression network named CARN is proposed to achieve automated quantitative measurement of the spine, which provides a reliable measurement for the clinical diagnosis and assessment of spinal diseases.
- The local structure-preserved manifold regularization (LSPMR) is proposed to generate discriminative feature embedding, which reduces the variability and largely improves the performance of multiple indices estimation.
- The adaptive local shape-constrained manifold regularization (ALSCMR) is proposed to alleviate overfitting. This provides a novel approach for multi-output regression to improve the generalization of the multi-output regression network.

In this work, we advance our preliminary attempt (Pang et al., 2018) on quantitative measurement of the spine in the following aspects:

- The LSPMR is proposed to obtain discriminative feature embedding, which largely reduces the variability in multi-output regression, and therefore achieves accurate multiple indices estimation.
- The robustness of the proposed CARN is validated by extended experiments using a larger dataset which contains 215 T1-weighted images and 20 T2-weighted images.
- The effectiveness of the proposed CARN is validated by comparing the performance with relevant machine learning based approaches.
- The loss weight of the local shape-constrained manifold regularization for each sample is determined adaptively. The sample with more reconstruction error of local linear representation in target output manifold has more probability to be an outlier and therefore has more loss weight of local shape-constrained manifold regularization to alleviate overfitting. As a result, the estimated indices are close to their real distribution.

2. Cascade amplifier regression network architecture

The proposed CARN architecture achieves automated multiple indices estimation of the spine through an expressive feature embedding obtained by the CAN and a linear regression model. As shown in Fig. 2, CAN is a network which provides an expressive

feature embedding by selective feature reuse using a series of AUs. The AU in CAN achieves selective feature reuse between the adjacent layers by a gate, multiplier, adder and concatenate operator. The selected feature map is generated by stimulating the effective feature map and suppressing the redundant feature map. Reusing these selected multi-level feature maps layer by layer, CAN provides an expressive feature embedding. The linear regression model in CARN is a fully connected network without nonlinear activation.

2.1. Mathematical formulation

The multiple indices estimation of the spine is summarized as a multi-output regression problem. Given a training dataset $T=\{x_i, y_i\}_{i=1}^{N}$, our goal is to train a multi-output regression model (i.e., the CARN) to learn the mapping f: $x \in R^{h \times w} \rightarrow y \in R^{d}$, where x_i and y_i denote the MR image and the corresponding multiple indices respectively, and N is the number of training samples. CARN aims to learn an expressive feature embedding and a reliable regressor simultaneously.

2.2. Cascade amplifier network for expressive feature embedding

The CAN achieves an expressive feature embedding by a combination of six AUs, two convolutional layers, five max-pooling layers, an average pooling layer, and a reshape operator as shown in Fig. 2 (a). AU is designed for selective feature reuse between adjacent layers. To achieve feature selection, the input feature map of AU is either stimulated or suppressed by an amplifier, whose amplification factor is learned automatically (details in Section 2.2.3). This feature selection procedure is based on the mechanism that the effective feature is stimulated while the redundant feature is suppressed. Whether the feature map is effective or not is determined by the amplification factor of the amplifier. The selected low-level feature is reused by concatenating the high-level feature. The selective feature reuse is achieved by CAN level by level; then the multi-level selective reused feature generates an expressive feature embedding. As shown in Fig. 2 (a), the first convolutional layer with a 7×7 kernel size and stride of 2 reduces the resolution of feature maps from 512×256 to 256×128, while the last convolutional layer with a 1×1 kernel size and stride of 1 linearly combines the feature maps for information integration. The max pooling with a 2×2 kernel size and a stride of 2 is used to provide translation invariance to the internal representation. The average pooling with a 4×2 kernel size and a stride of 4×2 is designed to reduce the dimensionality of feature maps, which finally are reshaped to a 1024 dimensional vector as the feature embedding.

2.2.1. Motivation of the amplifier unit

The architecture of AU is motivated by the gating mechanisms and skip connections. The gate controls the information propagation in networks and has proven beneficial to Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network and Gated Recurrent Unit (GRU) (Cho et al., 2014). LSTM achieved long-term memory for sequence information by input gate, output gate, and forget gate. GRU achieved a similar function of LSTM with a simpler structure composed of an update gate and reset gate. The gates in LSTM and GRU drop any information that is found to be irrelevant in the subsequent time step and carry any useful information from previous time step. The mechanism of the gates in LSTM and GRU is summarized as the selective feature reuse between different time steps. Skip connections have proven beneficial in several networks, such as DenseNet (Huang et al., 2017) and residual network (He et al., 2016), which demonstrate that reusing the feature maps of the previous layer in the network can improve the performance of the network. However, the reused feature maps of the previous layer without selection using skip connections may be redundant.

The AU is different from the LSTM and GRU although the AU is motivated by them. The tasks they solve are different. The AU is designed to process the single image while the LSTM and GRU are used to handle sequential data. Moreover, their objectives are different. The AU is designed to selectively reuse the information of the previous layer in the network while the LSTM and GRU aim to selectively reuse the information of the previous time step.

2.2.2. Amplifier unit architecture

Motivated by the gating mechanisms and skip connections, AU (as demonstrated in Fig. 2 (b)) is proposed to selectively reuse the features between the adjacent layer of the network. As the most crucial component of CAN, AU is composed of a gate for controlling the information propagation between adjacent layers, a convolutional layer with a 3×3 kernel size and stride of 1 for extracting a linear feature map, which is used to control the gate, a batch normalization layer with relu activation for producing nonlinear feature map, a multiplier, an adder, and a concatenation operator with batch normalization for combining the selected feature map and nonlinear feature map. The input t of AU goes through a convolutional layer which produces the linear feature map $f_l(t)$ for guiding feature selection. Then the $f_l(t)$ flows into two paths. One path consists of batch normalization and relu activation for generating nonlinear feature map $f_n(t)$. The other path is a gate composed of a convolutional layer and tanh activation, which generates output $f_g(t)$ of the gate for selecting feature map. The $f_g(t)$ flows into a multiplier followed by an adder, and generates the selected feature map $f_s(t)$. Finally, the $f_n(t)$ and $f_s(t)$ are concatenated along the channel axis and normalized by the batch normalization layer to generate the output feature map $f_{out}(t)$.

As mentioned above, the procedure of AU is mathematically described as follows:

$$f_l(t) = w_l * t + b_l \tag{1}$$

where w_l and b_l are the convolution kernel weight and bias of the convolutional layer respectively, and $*$ is the convolutional operator.

$$f_n(t) = relu(bn(f_l(t))) \tag{2}$$

where bn and $relu$ denote the batch normalization and relu activation respectively.

$$f_g(t) = tanh(bn(w_g * f_l(t) + b_g)) \tag{3}$$

where w_g and b_g are the convolution kernel weight and bias in the gate respectively.

$$f_s(t) = t \odot f_g(t) + t = t \odot (f_g(t) + 1) \tag{4}$$

where $\odot$ denotes the element-wise multiplication.

$$f_{out}(t) = bn(f_n(t) \oplus f_s(t)) \tag{5}$$

where $\oplus$ denotes the concatenation operator.

2.2.3. Feature selection mechanism

The selected feature map f_s is equivalent to be calculated by stimulating or suppressing the input feature map t via an amplifier. As demonstrated in Eq. 4, the value of each pixel in the selected feature map f_s is obtained by multiplying the corresponding value in the input feature map t with an amplification factor $[f_g(t) + 1]$, which ranges from 0 to 2. When the amplification factor is less than 1, the input feature map is suppressed. If the amplification factor is more than 1, the input feature map is stimulated. When the amplification factor is 1, the input feature map is directly propagated to the output like DenseNet (Huang et al., 2017). The amplification factor is determined by the temporary linear feature map $f_l(t)$, which is automatically learned during training.

2.3. Linear regression model for multiple indices estimation

The linear regression model, i.e., a fully connected layer, is capable of estimating the indices because the discriminative feature embedding reduces the variability of inter-subjects and simplifies the relationship between the latent feature space and target output manifold. Based on the expressive and discriminative feature embedding $h(x_i)$ (i.e., the output of the CAN), the output of the linear regression model is calculated as:

$$f(x_i) = w_o h(x_i) + b_o \tag{6}$$

where w_o and b_o are the weights matrix and bias of the linear regression respectively.

3. Loss function with manifold regularization

The loss function improves the spinal indices estimation accuracy by combining a preliminary loss $loss_p$ with LSPMR loss $loss_l$ in conjunction with ALSCMR loss $loss_a$. The preliminary loss is designed to minimize the distance between the estimation of indices and the ground truth. As shown in Fig. 3, the LSPMR is employed to achieve discriminative feature embedding by preserving the local geometrical structure of the latent feature space as same as the target output manifold. The ALSCMR is aimed at alleviating overfitting and generating realistic results by obliging the output of CARN to lie on the target output manifold using local linear representation. The total loss function is defined as follows:

$$loss_t(w) = \overbrace{loss_p(w)}^{\text{preliminary loss}} + \lambda_l \overbrace{loss_l(w)}^{\text{LSPMR loss}} + \lambda_a \overbrace{loss_a(w)}^{\text{ALSCMR loss}} \tag{7}$$

where λ_l and λ_a are scaling factors controlling the relative importance of the loss terms. The preliminary loss function is defined as follows:

$$loss_p(w) = \frac{1}{N \times d} \sum_{i=1}^{N} \|y_i - f(x_i)\|_1 + \lambda_p \sum_i \|w_i\|_2 \tag{8}$$

where the first term is the mean absolute error (MAE) of the regression model; the second term is the $l2$-norm regularization for the trainable weight w_i in CARN; λ_p is a hyper-parameter.

3.1. Local structure-preserved manifold regularization (LSPMR)

LSPMR is proposed to extract features associated with multiple indices for discriminative representation, which reduces the variability and simplifies the relationship between the latent feature space and target output manifold (Zhen et al., 2015). As shown in Fig. 3, constrained by LSPMR, the MR image is mapped to the latent feature space, which preserves the local geometrical structure of the target space. In contrast to the conventional manifold regularization (Zhen et al., 2017), the LSPMR combines manifold regularization and deep learning for discriminative feature embedding, which is rarely studied for multi-output regression.

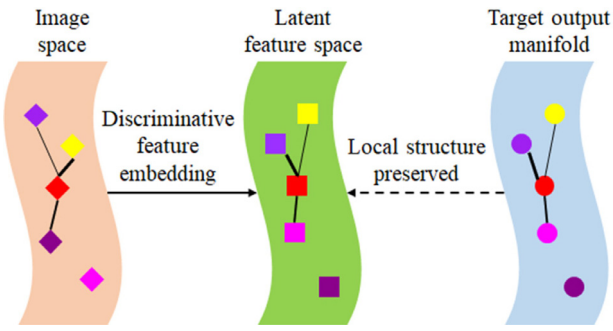


Fig. 3. The LSPMR achieves the discriminative feature embedding by preserving the local geometric structure of the latent feature space as same as the target output manifold. The points lying in the image space, latent feature space, and target output manifold with the same color are the same sample.

The LSPMR is based on the mechanism that the neighboring samples on the target output manifold are similar in the latent feature space. To this end, we construct a k-nearest neighbor graph $G = (V, E)$, where V and E represent vertices and edges between the vertices. The graph is constructed on the multivariate targets $\{y_i\}_{i=1}^N$ to represent the local neighbor relationship. $\hat{S} \in R^{N \times N}$ is denoted as the similarity matrix with non-negative elements corresponding to the edge weight of the graph G, where each element $\hat{S}_{i,j}$ is calculated by a heat kernel with parameter $\sigma_1 = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j>i}^{N} \|y_i - y_j\|_2$:

$$\hat{S}_{i,j} = \begin{cases} \exp\left(-\frac{\|y_i - y_j\|_2^2}{2\sigma_1^2}\right), y_j \in N_{k_1}(y_i), i \neq j, i, j = 1, 2, \ldots, N \\ 0, otherwise \end{cases} \tag{9}$$

where $\|\bullet\|_2$ denotes the $l2$-norm and $N_{k_1}(y_i)$ obtained by Euclidean distance represents the k_1-nearest neighbors of y_i. The $l2$-norm achieved excellent performance for finding k-nearest neighbors in several studies (Wu et al., 2016; Liu et al., 2010; Huang et al., 2014; Pang et al., 2017; 2015). Thus, we adopt the $l2$-norm, i.e., Euclidean distance, to obtain the k_1-nearest neighbors. Moreover, the $l2$-norm, i.e., Euclidean distance, provides a good approximation to geodesic distance for neighboring points in the manifold (Tenenbaum et al., 2000). Therefore, we use the $l2$-norm when constructing the similarity matrix in Eq. (9). The graph G using Eq. (9) as the edge weight is a directed graph and $\hat{S}$ is asymmetric. To simplify the computation, the asymmetric matrix $\hat{S}$ is replaced with a symmetric similarity matrix $S \in R^{N \times N}$ to construct an undirected graph G:

$$S = \max\left(\hat{S}, \hat{S}^T\right) \tag{10}$$

where $\max\left(\hat{S}, \hat{S}^T\right)$ denotes the element-wise maximum operation and $\hat{S}^T$ is the transpose of matrix $\hat{S}$. The local structure-preserved manifold regularization (LSPMR) loss is defined as follows:

$$loss_l = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{N} u(S_{i,j})} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{i,j} \|h(x_i) - h(x_j)\|_2^2 \tag{11}$$

$$u(t) = \begin{cases} 1, t > 0 \\ 0, t \leq 0 \end{cases} \tag{12}$$

The loss function of Eq. (11) is motivated by the idea of Laplacian Eigenmaps (Belkin and Niyogi, 2003), which incurs a penalty when similar vertexes are mapped far away in the latent feature space. As a result, the neighboring samples on the target output manifold are similar in the latent feature space. Note that we use $l2$-norm in Eq. (11) because the $l2$-norm is a good choice to construct Laplacian Eigenmaps (Zhong et al., 2016; Yang et al., 2018; Belkin and Niyogi, 2003), which can be optimized using stochastic gradient descent (Wang et al., 2016). The LSPMR loss is rephrased as follows:

$$\begin{aligned} loss_l &= \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{N} u(S_{i,j})} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{i,j} \|h(x_i) - h(x_j)\|_2^2 \\ &= \frac{2}{\sum_{i=1}^{N} \sum_{j=1}^{N} u(S_{i,j})} tr(H^T L H) \end{aligned} \tag{13}$$

where $L = D - S$, $D \in R^{N \times N}$ is a diagonal matrix, $D_{i,i} = \sum_{j=1}^{N} S_{i,j}$; H is the samples matrix in the latent feature space, $H_{i,\bullet} = h(x_i)$; $tr()$ denotes the trace of a matrix.

Note that the LSPMR is just a regularization term of the loss function, which doesn't change the architecture of the network.

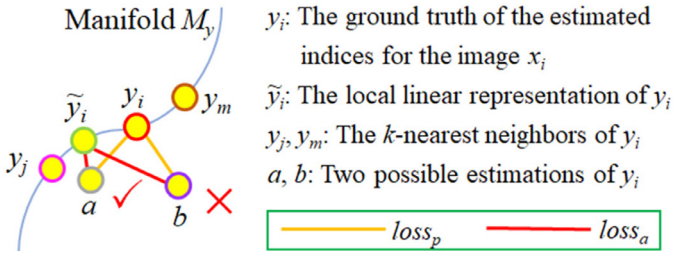


Fig. 4. ALSCMR ensures the indices estimation a is more consistent with their real distribution than b since $loss_p(a) + loss_a(a) < loss_p(b) + loss_a(b)$ with ALSCMR while $loss_p(a) = loss_p(b)$ without ALSCMR.

During the training phase, the CAN generates a discriminative feature embedding with the LSPMR. The trainable parameters in the CAN are learned. During the test phase, using the learned trainable parameters of the CAN, CAN maps the input MR image to the latent feature space.

3.2. Adaptive local shape-constrained manifold regularization (ALSCMR)

The estimated result of the indices of the spine is possible to be out of their real distribution. For instance, as shown in Fig. 4, y_i, y_j, and y_m are the target outputs of samples. The points a and b are two possible estimations of y_i. The distances between the two estimations (the points a and b) and the target output y_i are the same, ie., they have an identical preliminary loss. However, the loss of point a should be smaller than point b as a is much closer to the local shape of the target output space than b. Hence, a is a better estimation of y_i than b.

A realistic and robust multiple indices estimation of the spine must satisfy a constraint that the predicted indices should be close to the target output manifold M_y which is spanned by $\{y_i\}_{i=1}^N$. Developing a reliable metric to evaluate the distance between the predicted indices and the target output manifold is indispensable.

Local linear representation (Pang et al., 2017) is introduced to evaluate the distance between the predicted indices and the target output manifold. We assume that a sample on manifold M_y can be approximately represented as a linear combination of several nearest neighbors from M_y. This assumption was verified in many applications (Wu et al., 2016; Pang et al., 2017). A sample y_i on M_y can be locally linearly represented as:

$$\begin{aligned} &y_i = \sum_{j=1}^{k_2} y_j \alpha_j + \varepsilon_i \approx \sum_{j=1}^{k_2} y_j \alpha_j = \tilde{y}_i \\ &s.t. \|\varepsilon_i\| < \tau, \sum_{j=1}^{k_2} \alpha_j = 1, \alpha_j \geq 0, y_j \in N_{k_2}(y_i) \end{aligned} \tag{14}$$

where ε_i is the reconstruction error and τ is a small non-negative constant. $N_{k_2}(y_i)$ obtained by Euclidean distance denotes the k_2-nearest neighbors of y_i on M_y and α_j is the reconstruction coefficient, which is calculated by LAE (Liu et al., 2010). Both LAE and LLC (Wang et al., 2010) can be utilized for local linear representation (Pang et al., 2017). LAE uses the convex combination of several nearest neighbors of the sample to be its representation, which has high probability to lie on the manifold. This is why we adopt LAE to local linear representation instead of LLC. As shown in Fig. 4, $\tilde{y}_i$ is the local linear representation of y_i using its k-nearest neighbors (here k is equal to 2) y_j and y_m. The local linear representation of y_i reflects the local manifold shape. If the predicted indices is close to $\tilde{y}_i$, it will be near the manifold M_y.

The adaptive local shape-constrained manifold regularization (ALSCMR) is proposed to reduce the distance between the predicted indices and the target output manifold. The ALSCMR loss is defined as:

$$loss_a(w) = \frac{1}{N \times d} \sum_{i=1}^{N} \lambda_i \| f(x_i) - \tilde{y}_i \|_1 \tag{15}$$

where λ_i is the adaptive loss weight controlled by the reconstruction error of local linear representation on the target output manifold:

$$\lambda_i = 1 - \exp\left(-\frac{\|\varepsilon_i\|_2^2}{2\sigma_2^2} \right) \tag{16}$$

In Eq. 16, the sample, whose reconstruction error of local linear representation is large, lies far away from the manifold, thus it has a high probability to be an outlier and has a large loss weight. Using the $loss_a$, the prediction of y_i is restricted to being close to the manifold M_y, thus a more realistic result is obtained (e.g., the model generate the point a as the estimation of y_i instead of point b in Fig. 4).

4. Experiment configurations

4.1. Datasets

There are two datasets including: 1) The T1-weighted dataset which includes 215 subjects is collected from multi-center and different manufacturers using the parameters as follows: repetition time (TR) = 600 msec; echo time (TE) = 14 msec; flip angle (FA) = 90°. There are four clinical groups in the subjects, including 101 patients with lumbar disc herniation (LDH), 18 patients with intervertebral disc degeneration (IDD), 29 patients with lumbar spondylolisthesis (LS), and 67 normal subjects. The average age of the 215 subjects is 54 ± 15 years with 123 females and 92 males. For each subject, a midsagittal spine 1.5T T1-weighted MR image is manually selected by the physician to conduct quantitative measurement of the spine. 2) The T2-weighted dataset which is collected from 20 subjects including one normal subject, 6 patients with LDH, and 13 LS. The average age of the 20 subjects is 66 ± 15 years with 14 females and 6 males. For each subject, a midsagittal spine 1.5T T2-weighted MR image is manually selected by the physician to conduct quantitative measurement of the spine. The parameters of the acquired images are as follows: repetition time (TR) = 2500 msec; echo time (TE) = 109 msec; flip angle (FA) = 90°. For both two datasets, the pixel spacings range from 0.4688 mm/pixel to 0.7813 mm/pixel. Images are resampled to 0.4688 mm/pixel and the 30 estimated indices are manually annotated in this space by two physicians with more than 10-year experiences using ITK-SNAP tool (Yushkevich et al., 2006) and our tool developed by Matlab. The average of the manual annotations obtained by the two physicians is used as the ground truth. It costs 20 minutes to manually annotate an image by a physician.

All images undergo three preprocessing steps. First, two landmarks, i.e., the top-left corner of the L1 vertebral body and the bottom-left corner of the L5 vertebral body, are manually marked for each image to provide reference for ROI cropping, in which five vertebral bodies, including L1, L2, L3, L4 and L5, and the corresponding five discs under them are enclosed. Second, the intensity of the MR images is normalized by histogram matching. Finally, the cropped images are resized to 512×256 pixels. In the resized cropped images space, the pixel sizes of different subjects are diverse, therefore the ratio between the estimated indices and physical heights of the image is used to be the output of CARN, and the estimated indices are obtained by multiplying the output of CARN with the physical heights of the image. Stratified 5-fold cross-validation is used to divide the training and test datasets. Specifically, the T1(T2) dataset is randomly divided into 5 folds and the percentage of samples for LDH, IDD, LS, and normal subjects in each fold is approximately equal. Tables 1 and 2 illustrate

Table 1
The number of subjects for each pathological class in each fold is approximately equivalent in the T1 dataset.

Pathological class	fold #1	fold #2	fold #3	fold #4	fold #5	Total
LDH	21	20	20	20	20	101
IDD	3	3	4	4	4	18
LS	6	6	5	6	6	29
Normal	13	14	14	13	13	67

Table 2
The number of subjects for each pathological class in each fold is approximately equivalent in the T2 dataset.

Pathological class	fold #1	fold #2	fold #3	fold #4	fold #5	Total
LDH	1	2	1	1	1	6
LS	3	2	3	3	2	13
Normal	0	0	0	0	1	1

the pathological statistical information for each fold. As shown in Tables 1 and 2, the number of subjects for each pathological class in each fold is approximately equal. Four folds are employed to train the model and the last fold is used for test. Since the deep learning methods like our CARN typically require large amounts of training data, the training data is augmented to improve the generalization performance by randomly rotating and cropping the images.

This study was approved by the Research Ethics Board of Western University (REBID: 17656E).

4.2. Training

All the networks are trained using stochastic gradient descent (SGD) with a weight decay of 0.005 and a Nesterov momentum (Sutskever et al., 2013) of 0.9 without dampening. The networks are based on Tensorflow and codes[1] run with a Tesla P100-SXM2 GPU. For the T1 dataset, the models are trained for 500 epochs with a batch size of 8. The learning rate is set to 0.04 initially and is lowered by 10 times at epoch 150 and 400. All parameters are empirically set as: $\lambda_p = 0.005$, $\lambda_l = 0.005$, $\lambda_a = 1$, $k_1 = 5$, $k_2 = 20$, $\sigma_2 = \frac{1}{2N}\sum_{i=1}^{N}\|\varepsilon_i\|_2$.

For the T2 dataset, three models are trained for the proposed CARN including:

- Model trained from scratch: The model are trained from scratch using the T2 dataset for 5000 epochs with a batch size of 8. The learning rate is set to 0.04 initially and is lowered by 10 times at epoch 1500 and 4000. All parameters are empirically set as: $\lambda_p = 0.005$, $\lambda_l = 0.005$, $\lambda_a = 1$, $k_1 = 5$, $k_2 = 5$, $\sigma_2 = \frac{1}{2N}\sum_{i=1}^{N}\|\varepsilon_i\|_2$.
- Pretrained model: The model are trained using the T1 dataset without fine-tuning on the T2 dataset.
- Fine-tuned model: The model are pretrained using the T1 dataset and fine-tuned on the T2 dataset. The pretrained model are fine-tuned for 150 epochs with a batch size of 8 and a learning rate of 0.004. All parameters are empirically set as: $\lambda_p = 0.005$, $\lambda_l = 0.005$, $\lambda_a = 1$, $k_1 = 5$, $k_2 = 5$, $\sigma_2 = \frac{1}{2N}\sum_{i=1}^{N}\|\varepsilon_i\|_2$.

4.3. Evaluation metric

The accuracy of indices estimation is evaluated by the mean absolute error (MAE) defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|a_i - b_i| \tag{17}$$

where a_i and b_i denote the predicted indice and the ground truth respectively, and N is the number of subjects.

5. Results and analysis

The performance of the proposed method is evaluated on T1 dataset and T2 dataset separately due to the variation between the T1-weighted and T2-weighted MR images.

5.1. Results on the T1 dataset

A set of experiments are presented to evaluate the performance of the proposed method on the T1 dataset including: 1) Overall performance; 2) effectiveness of CAN; 3) effectiveness of local structure-preserved manifold regularization; 4) effectiveness of adaptive local shape-constrained manifold regularization; 5) Performance comparison with relevant methods.

5.1.1. Overall performance

As shown in Fig. 5, the proposed CARN achieves accurate multiple indices estimation for the spine, which can be attributed to the expressive feature embedding, discriminative feature embedding, and the real distribution learning of estimated indices. Specifically, CARN achieves average MAE of 1.23 ± 1.04 mm, 1.22 ± 1.05 mm, and 1.22 ± 1.04 mm for IDHs, VBHs, and total indices estimation with respect to their manually obtained ground truth values. These errors are small referring to the maximums of IDHs (20.92 mm) and VBHs (36.71 mm) in the dataset. The performance of the CARN for the normal subjects is better than the patients due to the variety of the patients' spine. Since the model tends to be overfitting which arises from less IDD subjects, the CARN achieves worse performance for IDD subjects than other subjects.

For visualization, Fig. 6 shows the results of the automated quantitative measurement of the spine for four typical T1-weighted images from LDH, IDD, LS, and normal subjects. As shown in Fig. 6, for the LDH subject, the intervertebral disc of L5/S1 herniates posteriorly. For the IDD subject, the IDHs of L5/S1 decrease due to intervertebral disc degeneration. For the LS subject, the vertebrae of L4 moves forward. In spite of these pathological changes, the CARN achieves accurate measurement for most of the indices.

5.1.2. Effectiveness of CAN

Owing to the CAN, the MAE relatively significantly decreased by 5.38%, 6.15%, and 6.15% for IDHs, VBHs, and total indices estimation respectively as shown in Table 3. The non-CAN is used to denote the method, in which AU is replaced with a traditional convolutional layer and the output feature channels are the same as

[1] Released code: https://github.com/pangshumao/CARN.

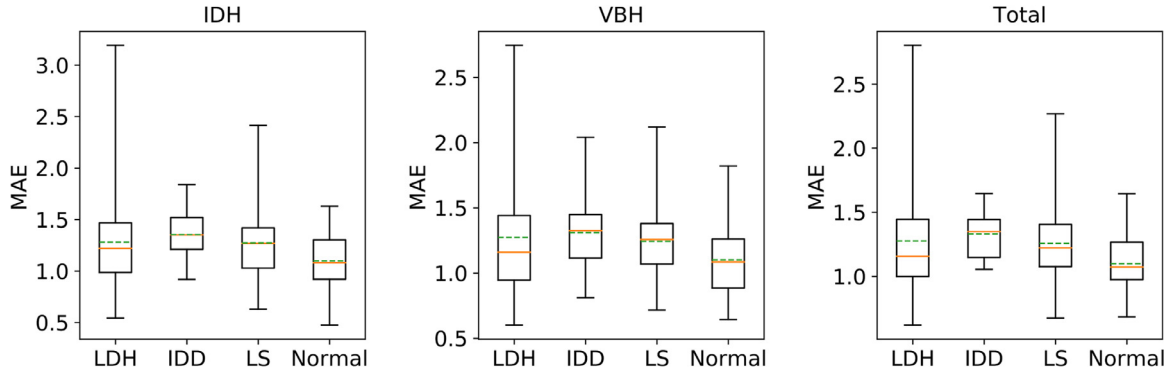


Fig. 5. The proposed CARN achieves accurate multiple indices estimation of the spine for normal subjects and patients of lumbar disc herniation (LDH), intervertebral disc degeneration (IDD), and lumbar spondylolisthesis (LS). The green dashed line denotes the mean of MAE; the orange solid line denotes the median of MAE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The proposed method significantly achieves lower MAE than the non-CAN in IDH (mm), VBH (mm), and total indices (mm) estimation on the T1 dataset. The * denotes $p < 0.01$ and ** denotes $p < 0.001$ for the paired t-test between the CAN approach and non-CAN method.

Method		IDH	VBH	Total
non-CAN		1.30 ± 1.08	1.30 ± 1.08	1.30 ± 1.08
CAN	4-AU	1.31 ± 1.10	1.37 ± 1.15**	1.34 ± 1.13*
	5-AU	1.28 ± 1.09	1.29 ± 1.10	1.28 ± 1.09
	6-AU	**1.23 ± 1.04****	**1.22 ± 1.05****	**1.22 ± 1.04****
	7-AU	1.23 ± 1.05**	1.25 ± 1.07**	1.24 ± 1.06**

Table 4
The proposed method significantly achieves lower MAE than the non-LSPMR in IDH (mm), VBH (mm), and total indices (mm) estimation on the T1 dataset. The ** denotes $p < 0.001$ for the paired t-test between the proposed approach and non-LSPMR method.

Method	IDH	VBH	Total
non-LSPMR	1.27 ± 1.07	1.26 ± 1.08	1.27 ± 1.07
Proposed	**1.23 ± 1.04****	**1.22 ± 1.05****	**1.22 ± 1.04****

the AU. To be fair, we train the non-CAN with the optimal hyper-parameters $\lambda_a = 1$ and $\lambda_l = 0.001$. As shown in Table 3, the effectiveness of CAN for the VBHs estimation is more significant than IDHs, which results from the fact that the CAN generates expressive feature embedding although pathological changes in the disc can reduce the intensity of the adjacent VB and lead to ambiguity in the boundary.

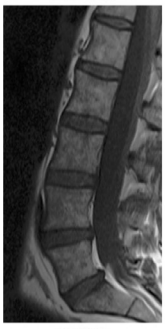
The AU in CAN stimulates the effective feature and suppresses the redundant feature during propagating feature map between adjacent layers. As shown in Figs. 7 and 2, t denotes an input feature map of the AU1; f_g represents an output feature map of the gate in the AU1; f_s is a selected feature map of the AU1; the right feature map $f_s - t$ is the difference between f_s and t. If $f_s - t > 0$, the feature map t is stimulated, otherwise, t is suppressed. The feature maps in the regions of abdominal cavity (the blue arrow regions) and ligamentum flavum (the green arrow region) are suppressed because these information are redundant for the indices estimation of the spine, while the feature maps in the regions of vertebral bodies (the green dashed boxes) are stimulated due to their effectiveness for the indices estimation of the spine. The selected feature map f_s is reused in the posterior layers. Thus an expressive feature embedding is obtained by the CAN. Our CAN effectively captures the structures of spinal images, as shown in Fig. 8. The layers AU1, AU2, and AU3 capture low-level visual features of spinal images, including the appearance of the vertebral bodies and intervertebral discs, textures of the vertebral bodies and intervertebral discs as well as the textures in the background. The layers of AU4 and AU5 extract more complex spinal structures by stimulating the features of these areas including vertebral bodies and intervertebral discs. The AU6 feature maps, with a resolution of 8×4, extract the indices-related features in each local area from the structures of all AU5 feature maps during the training procedure.

Influence of the depth of CAN. The number of AU is set to 4, 5, 6, and 7 (denoted as 4-AU, 5-AU, 6-AU, and 7-AU respectively) to investigate the influence of the depth of the network. For the 4-AU network, the last two max pooling layer and last two AU in Fig. 2 are removed. For the 5-AU network, the last max pooling layer and last AU in Fig. 2(a) are removed. For the 7-AU network, a max pooling layer and AU are added to the last AU of the Fig. 2(a). As shown in Table 3, the 6-AU network achieves best performance. Too shallow networks (4-AU and 5-AU) are incapable of obtaining effective feature embedding and feature embedding obtained by seven AUs is too abstract. Therefore, the number of the AU is set to 6.

5.1.3. Effectiveness of local structure-preserved manifold regularization

The results indicate that the LSPMR improves the performance of indices estimation for the spine by enhancing the discrimination of feature embedding. Table 4 compares our method against the same network using the loss function without LSPMR (non-LSPMR), i.e., the combination of $loss_p$ with $loss_a$. Using the LSPMR, the MAE relatively significantly decreased by 3.15%, 3.17%, and 3.94% for IDHs, VBHs, and total indices estimation respectively. As shown in Fig. 9, the training LSPMR loss of the proposed method is lower than the non-LSPMR. Using the LSPMR, the images with similar estimated indices lie in a closed region in the latent feature space. Thus the discriminative feature embedding is obtained and the performance of indices estimation for the spine is improved.

The LSPMR improves the discrimination of feature embedding which is demonstrated by applying t-SNE (Maaten and Hinton, 2008) to the feature embedding inferred from the 215 spine images. As a result, each feature embedding of a spine image is mapped to a two-dimensional vector. Then we can visualize each vector as a point on a two-dimensional space. The spine images are classified into three categories by the k-means cluster according to the ground truth of the estimated indices. As shown in Fig. 10, each point denotes a spine image and different categories are labeled as different colors. For the proposed method, the points of the same category are near from each other. However, for the non-LSPMR, the points belonging to different categories are mixed with each other. This further validates that the LSPMR is able to produce



prediction Ground Truth

LDH

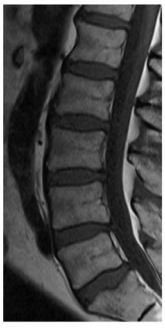
	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	25.85	28.62	30.14	30.94	31.86	6.86	8.78	8.42	11.15	12.69
	26.21	29.07	30.48	32.28	30.86	7.06	8.73	7.91	11.15	11.50
Middle	24.92	24.24	24.80	25.56	24.63	9.32	12.78	12.03	11.22	10.31
	24.51	24.93	25.52	27.45	24.81	9.10	11.48	11.48	9.94	10.41
Posterior	26.95	27.44	27.73	27.12	22.14	6.78	7.70	6.98	7.06	7.02
	26.58	28.17	28.90	26.79	22.59	5.53	7.94	7.26	6.86	6.29

IDD

	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	24.70	26.31	27.76	28.04	25.57	5.56	7.59	9.34	14.55	11.83
	23.92	26.64	28.59	28.24	26.03	7.26	4.66	9.98	16.72	9.94
Middle	25.45	26.03	24.88	24.66	22.84	7.12	7.02	11.26	13.40	8.79
	26.40	25.00	25.88	24.66	25.02	8.42	7.38	11.96	13.79	6.85
Posterior	27.33	28.44	28.26	26.67	24.27	5.30	5.27	7.16	7.63	6.65
	29.41	29.08	29.56	28.90	25.40	5.24	4.61	6.50	7.06	6.17

LS

	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	22.34	22.76	24.79	27.54	25.44	7.11	8.27	9.19	9.17	9.38
	21.88	22.43	27.40	27.66	24.08	7.48	6.17	9.27	10.44	10.75
Middle	22.37	23.11	24.14	23.29	22.85	5.98	5.23	8.47	7.57	3.28
	23.37	23.20	23.97	26.67	23.61	6.16	4.94	6.86	8.51	3.10
Posterior	24.30	24.53	24.21	23.12	21.22	5.10	4.03	4.64	6.39	3.66
	22.73	23.72	25.72	23.59	20.85	6.25	3.06	4.50	6.58	3.42



Normal

	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	27.36	29.11	31.69	31.25	31.09	10.65	10.17	12.08	13.93	15.32
	28.75	28.65	32.41	30.57	30.71	10.48	9.64	10.81	14.08	15.57
Middle	24.74	25.09	26.04	25.94	26.61	11.93	12.43	13.41	11.80	11.07
	25.66	25.73	25.08	26.80	26.50	11.71	12.56	13.36	12.03	11.78
Posterior	28.68	27.71	27.75	26.97	25.48	8.35	8.68	8.23	7.42	6.76
	26.80	28.82	24.77	28.77	24.10	9.80	8.39	9.61	7.22	6.85

Fig. 6. Visualization of the automated quantitative measurement of the spine for four typical T1-weighted images from LDH, IDD, LS, and normal subjects. The proposed CARN achieves accurate prediction (Unit: mm) of most of VBH and IDH for different pathological subjects.

discriminative feature embedding by preserving the local structure between the latent feature space and the target output manifold.

5.1.4. Effectiveness of adaptive local shape-constrained manifold regularization

The results show that the ALSCMR improves the generalization of the network for indices estimation of the spine. Table 5 compares the proposed method against the same network using the loss function without ALSCMR (non-ALSCMR), i.e., the combination of $loss_p$ and $loss_l$. Using the ALSCMR, the performance relatively significantly improved by 2.38%, 1.61%, and 2.40% for IDHs, VBHs, and total indices estimation respectively. Compared to non-ALSCMR, the proposed method achieves high training MAE (0.19 mm vs 0.17 mm, 0.20 mm vs 0.17 mm) but low test MAE (1.23 mm vs 1.26 mm, 1.22 mm vs 1.24 mm) for IDH and VBH estimations respectively. This results from the fact that ALSCMR restricts the output of the CARN to the target output manifold and the distribution of the estimated indices are close to their real

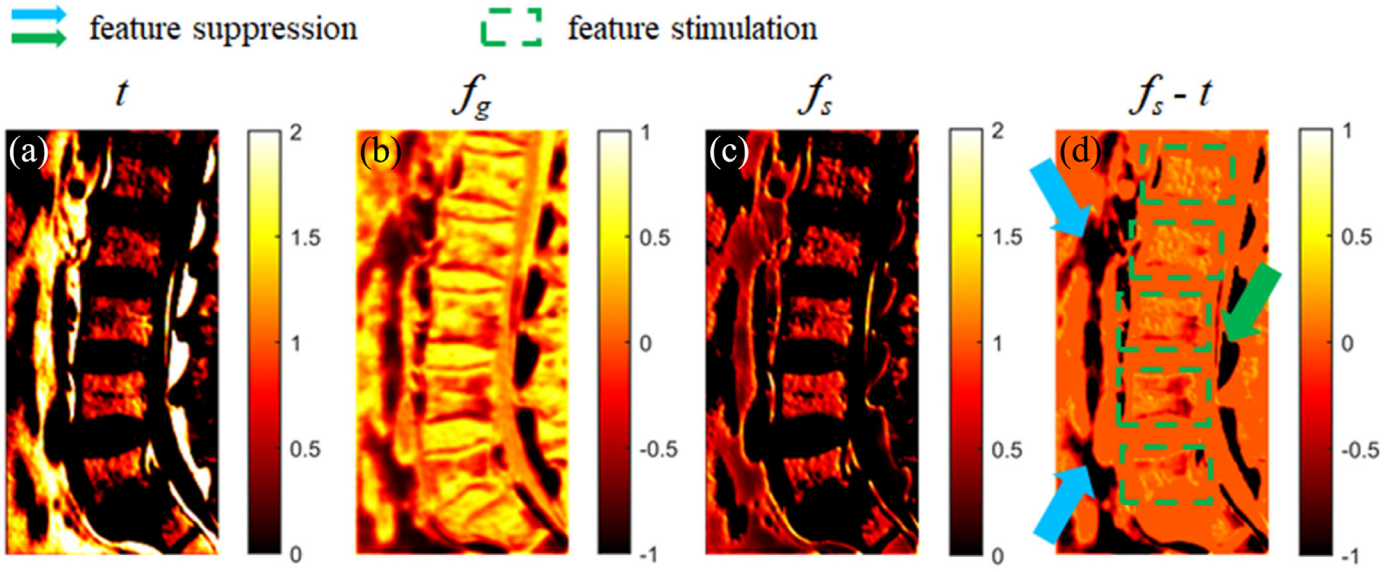


Fig. 7. The expressive feature embedding is obtained by stimulating the effective feature and suppressing the redundant feature in AU. (a) The input feature map of the AU1; (b) the output feature map of the gate in the AU1; (c) the selected feature map of the AU1; (d) the difference between (c) and (a).

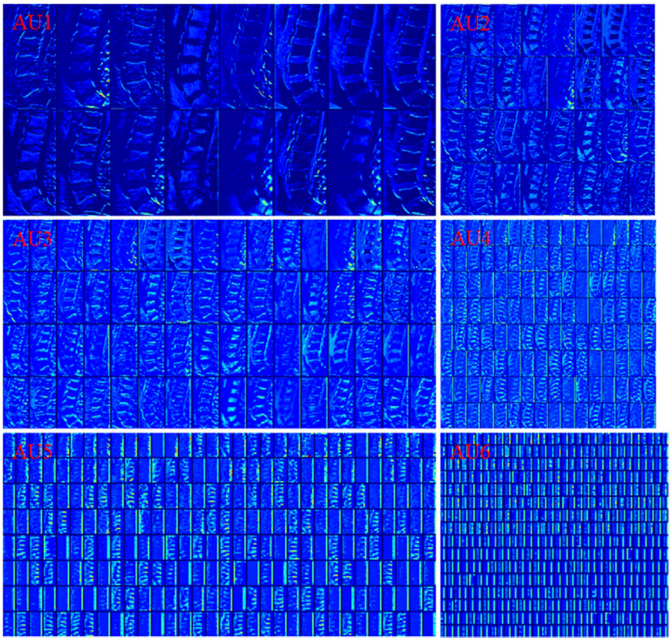


Fig. 8. Visualization of feature maps obtained by our CAN for an example spinal image.

Table 5
The proposed method achieves higher training MAE but lower test MAE than the non-ALSCMR in IDH (mm), VBH (mm), and total indices (mm) estimation on the T1 dataset. The * denotes $p < 0.01$ for the paired t-test between the proposed approach and non-ALSCMR method.

Method		non-ALSCMR	Proposed
IDH	training	0.17 ± 0.15	0.19 ± 0.23*
	test	1.26 ± 1.06	**1.23 ± 1.04***
VBH	training	0.17 ± 0.14	0.20 ± 0.29*
	test	1.24 ± 1.06	**1.22 ± 1.05***
Total	training	0.17 ± 0.14	0.20 ± 0.26*
	test	1.25 ± 1.06	**1.22 ± 1.04***

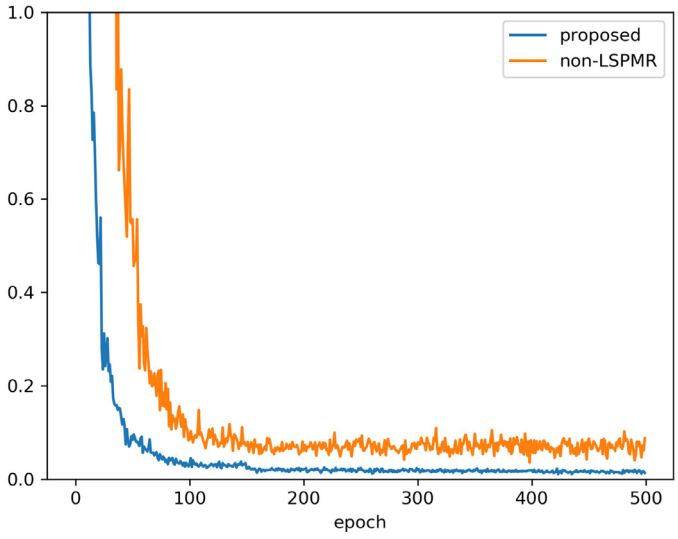


Fig. 9. The proposed method has lower training LSPMR loss thus generates more discriminative feature embedding compared to the non-LSPMR.

distribution, which reduces the impact of outliers and improves the generalization of the network. The loss weight of the ALSCMR for each sample is determined adaptively. The sample with more reconstruction error of local linear representation in output manifold has more probability to be an outlier and therefore has more loss weight of the ALSCMR to alleviate overfitting.

5.1.5. Performance comparison with relevant methods

The proposed CARN achieves best performance in indices estimation of the spine compared to the relevant machine learning based methods including MCDBN+RF (Zhen et al., 2016), Multi-features+RF (Zhen et al., 2014), HOG+AKRF (Hara and Chellappa, 2014), and HOG+SSVR (Sun et al., 2017), and deep learning based approaches including DenseNet (Huang et al., 2017), and CARN-LSCMR (Pang et al., 2018). All relevant approaches are implemented using the optimal parameters. The DenseNet are configured as follows:

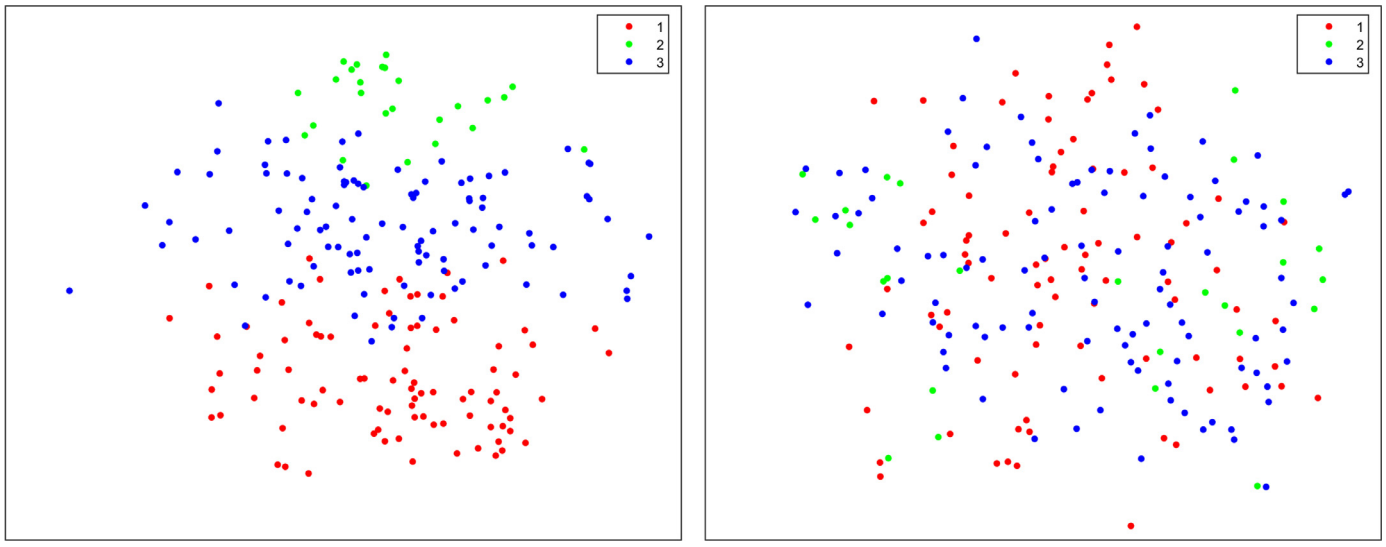


Fig. 10. The LSPMR (left) is able to learn more discriminative feature embedding compared to non-LSPMR (right) as shown by the t-SNE representation of feature embeddings. Points are colored according to the categories produced by k-means cluster for the ground truth of the estimated indices.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
The proposed method achieves lowest MAE in IDH (mm), VBH (mm), and total indices (mm) estimation on the T1 dataset compared to relevant machine learning based methods and deep learning based approaches. The * denotes $p < 0.01$ and ** denotes $p < 0.001$ for the paired t-test between the proposed approach and the relevant method.

Method	IDH	VBH	Total
MCDBN+RF	1.55 ± 1.29**	1.45 ± 1.19**	1.50 ± 1.24**
Multi-features+RF	1.50 ± 1.23**	1.42 ± 1.20**	1.46 ± 1.21**
HOG+AKRF	1.54 ± 1.33**	1.48 ± 1.22**	1.51 ± 1.28**
HOG+SSVR	1.42 ± 1.18**	1.36 ± 1.14**	1.39 ± 1.16**
DenseNet-14	1.32 ± 1.12**	1.28 ± 1.11**	1.30 ± 1.12**
DenseNet-18	1.30 ± 1.12**	1.27 ± 1.09**	1.29 ± 1.11**
DenseNet-22	1.28 ± 1.10**	1.26 ± 1.08*	1.27 ± 1.09*
DenseNet-26	1.29 ± 1.10**	1.28 ± 1.12**	1.28 ± 1.11**
CARN-LSCMR	1.28 ± 1.06**	1.27 ± 1.07*	1.27 ± 1.07*
Proposed	**1.23 ± 1.04**	**1.22 ± 1.05**	**1.22 ± 1.04**

- DenseNet-14: In this version of DenseNet, the resolution of the feature maps is reduced from 512×256 to 64×32 using two convolutional layers with a 7×7 kernel size and stride of 2 and a 3×3 max pooling with a stride of 2. Then four dense blocks and three transition layers are followed. Each dense block consists of two 3×3 convolutional layers with a stride of 1. The transition layer includes a 1×1 convolutional layer with a stride of 1 and a followed 2×2 average pooling layer with a stride of 2. Note that before each convolutional layer, batch normalization and relu are included. Finally, a global average pooling layer is utilized to obtain the feature embedding and a linear fully-connected layer generates the estimated indices. The growth rate is 48 and the depth of the network is 14.
- DenseNet-18: This network is the same with DenseNet-14 except that the number of the convolutional layer in each dense block is 3 and the depth of the network is 18.
- DenseNet-22: This network is the same with DenseNet-14 except that the number of the convolutional layer in each dense block is 4 and the depth of the network is 22.
- DenseNet-26: This network is the same with DenseNet-14 except that the number of the convolutional layer in each dense block is 5 and the depth of the network is 26.

As shown in Table 6, the proposed method significantly improves the accuracy by 13.38%, 10.29%, and 12.23% for IDHs, VBHs, and total indices estimation respectively compared to HOG+SSVR which achieves the best performance among the traditional machine learning approaches. This is owed to the fact that the proposed method can generate expressive and discriminative feature embedding by selective feature reuse and LSPMR while the handcrafted feature extraction approach HOG is incapable of capturing the task-specific feature. Among the DenseNet approaches, DenseNet-22 achieves impressive performance, which however is inferior to the proposed method. Even though DenseNet reuses feature maps in each block, it doesn't obtain expressive feature embedding due to the lack of feature selection. Furthermore, the MAE of the proposed method is significantly lower than the CARN-LSCMR by 3.91%, 3.94%, and 3.94% for IDHs, VBHs, and total indices estimation respectively. This validates the effectiveness of LSPMR and ALSCMR.

Table 7
The proposed fine-tuned model achieves impressive performance in terms of MAE in IDH (mm), VBH (mm), and total indices (mm) estimation on the T2 dataset.

Method	IDH	VBH	Total
Trained from scratch	1.35 ± 1.12	1.39 ± 1.19	1.37 ± 1.15
Pretrained	1.34 ± 1.07	1.49 ± 1.29	1.41 ± 1.19
Fine-tuned	**1.16 ± 0.93**	**1.32 ± 1.19**	**1.24 ± 1.07**

5.2. Results on the T2 dataset

The proposed CARN is robust for the variation of the images intensity. As shown in Table 7, the model pretrained on the T1 dataset achieves promising performance for the T2 dataset. This demonstrates that the generalization of the CARN is robust. The model fine-tuned on the T2 dataset achieves the best performance compared to the model trained from scratch because the T2 dataset is small and the model trained from scratch tends to overfit. The T1 and T2 images share common features such as edge and shape information although their intensity distribution is various. Therefore, the pretrained model using T1 dataset can facilitate the fine-tuned model to achieve a better performance.

For visualization, Fig. 11 shows the results of the automated quantitative measurement of the spine for three typical T2-weighted images from LDH, LS, and normal subjects. Results show

LDH

prediction Ground Truth

	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	24.63	27.44	27.33	28.58	25.66	8.96	9.97	11.62	11.58	14.04
	25.68	27.16	27.10	27.53	27.41	7.82	10.08	12.67	9.93	13.64
Middle	22.43	23.88	22.97	22.52	21.66	8.73	9.72	10.94	9.39	10.47
	23.42	23.31	22.96	23.00	22.36	8.03	9.41	10.75	8.46	8.93
Posterior	26.36	27.48	26.31	23.60	23.23	5.31	6.25	6.92	8.26	6.92
	27.16	26.80	26.39	23.09	22.98	5.53	5.91	6.17	7.35	6.75

LS

	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	23.04	25.56	26.14	27.41	25.92	7.68	9.14	13.07	13.41	15.73
	23.40	24.42	24.73	28.09	28.17	8.59	5.32	13.75	12.19	18.29
Middle	20.46	21.83	21.92	22.75	20.62	7.86	8.52	10.04	10.59	10.64
	19.94	21.69	22.31	22.62	21.92	8.66	6.40	10.30	10.10	12.33
Posterior	24.55	26.85	25.31	23.11	22.40	5.44	4.74	6.04	7.59	7.54
	24.75	26.85	24.59	22.52	22.86	5.53	3.90	6.74	8.02	8.28

Normal

	VBH					IDH				
	L1	L2	L3	L4	L5	L1/L2	L2/L3	L3/L4	L4/L5	L5/S1
Anterior	25.43	29.77	29.43	28.70	26.85	8.77	11.17	13.33	12.50	16.84
	27.00	27.24	26.97	26.45	25.01	9.80	14.30	14.45	14.78	14.95
Middle	23.95	25.50	24.59	24.94	22.93	9.96	10.90	11.56	10.70	11.28
	24.04	24.51	24.46	24.33	26.01	10.29	12.79	11.26	10.94	10.37
Posterior	28.98	28.56	28.61	25.45	24.27	7.06	7.48	8.17	8.78	7.15
	28.38	27.99	28.11	26.04	22.70	7.91	10.02	7.54	9.35	9.19

Fig. 11. Visualization of the automated quantitative measurement of the spine for three typical T2-weighted images from LDH, LS, and normal subjects. The proposed fine-turned CARN achieves accurate prediction (Unit: mm) of most of VBH and IDH for different pathological subjects.

that the proposed fine-turned CARN achieves accurate prediction of most of VBH and IDH for different pathological subjects. For the LDH subject in Fig. 11, the fine-turned CARN achieves accurate indices estimation for the posterior IDH of L5/S1 and other most of indices even though the intervertebral disc of L5/S1 herniates posteriorly. For the LS subject, the fine-turned CARN achieves accurate indices estimation for the IDHs of L3/L4 and L4/L5 even though the vertebrae of L4 moves forward. The accuracy of the automated quantitative measurement of the spine for the normal subject is inferior to other pathological subjects because there are only one normal subject in the T2 dataset.

6. Conclusion

We have presented an accurate and robust method for automated quantitative measurement of the spine using CARN with manifold regularization. The CAN achieves expressive feature embedding by reusing the selected feature. The feature selection is implemented by stimulating the effective feature but suppressing the redundant feature during propagating feature map between adjacent layers. Whether the feature is effective or redundant is automatically learned during training. The LSPMR enhances the discrimination of feature embedding by preserving the local geometric structure between the latent feature space and target output manifold invariant. Using the expressive and discriminative feature embedding in conjunction with the ALSCMR for alleviating overfitting, the proposed approach achieves accurate and robust estimation for indices of the spine.

Although the proposed CARN achieves accurate and robust automated quantitative measurement of the spine, it suffers from several limitations. For example, the image to be processed must be cropped to include only the lumbar vertebrae. The effect of extra, partial, fused, and missing vertebrae to the performance of the network is unknown. We will overcome these limitations in our future work.

Acknowledgments

Computations were performed using the data analytics Cloud at SHARCNET (http://www.sharcnet.ca) provided through the Southern Ontario Smart Computing Innovation Platform (SOSCIP); the SOSCIP consortium is funded bythe Ontario Government and the Federal Economic Development Agency for Southern Ontario. Financial support for this work was partly provided by the China Scholarship Council (no. 201708440350), the National Natural Science Foundation of China (no. U1501256), and the Science and Technology Project of Guangdong Province (no. 2015B010131011). No other potential conflict of interest relevant to this article was reported.

References

Barbieri, P.D., Pedrosa, G.V., Traina, A.J.M., Nogueira-Barbosa, M.H., 2015. Vertebral body segmentation of spine MR images using superpixels. In: Computer-Based Medical Systems (CBMS), 2015 IEEE 28th International Symposium on. IEEE, pp. 44–49.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15 (6), 1373–1396. doi:10.1162/089976603321780317.

Brinckmann, P., Grootenboer, H., 1991. Change of disc height, radial disc bulge, and intradiscal pressure from discectomy an in vitro investigation on human lumbar discs. Spine 16 (6), 641–646. doi:10.1097/00007632-199106000-00008.

Castro, I., Humbert, L., Whitmarsh, T., Lazary, A., Barquero, L.D.R., Frangi, A.F., 2012. 3d reconstruction of intervertebral discs from t1-weighted magnetic resonance images. In: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on. IEEE, pp. 1695–1698.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics doi:10.3115/v1/d14-1179.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, pp. 886–893 vol. 1. doi:10.1109/CVPR.2005.177.

Hara, K., Chellappa, R., 2014. Growing Regression Forests by Classification: Applications to Object Pose Estimation. In: Computer Vision – ECCV 2014. Springer International Publishing, pp. 552–567. doi:10.1007/978-3-319-10605-2_36.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE doi:10.1109/cvpr.2016.90.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269.

Huang, M., Yang, W., Wu, Y., Jiang, J., Chen, W., Feng, Q., 2014. Brain tumor segmentation based on local independent projection-based classification. IEEE Trans. Biomed. Eng. 61 (10), 2633–2645.

Jarman, J.P., Arpinar, V.E., Baruah, D., Klein, A.P., Maiman, D.J., Muftuler, L.T., 2014. Intervertebral disc height loss demonstrates the threshold of major pathological changes during degeneration. Eur. Spine J. 24 (9), 1944–1950. doi:10.1007/s00586-014-3564-8.

Korez, R., Ibragimov, B., Likar, B., Pernuš, F., Vrtovec, T., 2017. Intervertebral disc segmentation in mr images with 3d convolutional networks. In: Medical Imaging 2017: Image Processing, 10133. International Society for Optics and Photonics, p. 1013306.

Liu, W., He, J., Chang, S.-F., 2010. Large graph construction for scalable semi-supervised learning. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 679–686.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. J.Mach.Learn.Res. 9 (Nov), 2579–2605.

McCloskey, E., Johansson, H., Oden, A., Kanis, J.A., 2012. Fracture risk assessment. Clin. Biochem. 45 (12), 887–893. doi:10.1016/j.clinbiochem.2012.05.001.

Pang, S., Jiang, J., Lu, Z., Li, X., Yang, W., Huang, M., Zhang, Y., Feng, Y., Huang, W., Feng, Q., 2017. Hippocampus segmentation based on local linear mapping. Sci. Rep. 7, 45501. doi:10.1038/srep45501.

Pang, S., Leung, S., Ben Nachum, I., Feng, Q., Li, S., 2018. Direct automated quantitative measurement of spine via cascade amplifier regression network. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 940–948.

Pang, S., Lu, Z., Yang, W., Wu, Y., Lu, Z., Zhong, L., Feng, Q., 2015. Hippocampus segmentation through distance field fusion. In: International Workshop on Patch-based Techniques in Medical Imaging. Springer, pp. 104–111.

Salamat, S., Hutchings, J., Kwong, C., Magnussen, J., Hancock, M.J., 2016. The relationship between quantitative measures of disc height and disc signal intensity with pfirrmann score of disc degeneration. SpringerPlus 5 (1). doi:10.1186/s40064-016-2542-5.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Sun, H., Zhen, X., Bailey, C., Rasoulinejad, P., Yin, Y., Li, S., 2017. Direct Estimation of Spinal Cobb Angles by Structured Multi-output Regression. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 529–540. doi:10.1007/978-3-319-59050-9_42.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: International conference on machine learning, pp. 1139–1147.

Tatoń, G., Rokita, E., Korkosz, M., Wróbel, A., 2014. The ratio of anterior and posterior vertebral heights reinforces the utility of DXA in assessment of vertebrae strength. Calcif. Tissue Int. 95 (2), 112–121. doi:10.1007/s00223-014-9868-1.

Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Tunset, A., Kjaer, P., Chreiteh, S.S., Jensen, T.S., 2013. A method for quantitative measurement of lumbar intervertebral disc structures: an intra- and inter-rater agreement and reliability study. Chiropr. Manual Ther. 21 (1), 26. doi:10.1186/2045-709x-21-26.

Videman, T., Battié, M.C., Gibbons, L.E., Gill, K., 2014. Aging changes in lumbar discs and vertebrae and their interaction: a 15-year follow-up study. Spine J. 14 (3), 469–478. doi:10.1016/j.spinee.2013.11.018.

Wang, C., Forsberg, D., 2016. Segmentation of Intervertebral Discs in 3D MRI Data Using Multi-atlas Based Registration. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 107–116. doi:10.1007/978-3-319-41827-8_10.

Wang, D., Cui, P., Zhu, W., 2016. Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1225–1234.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 3360–3367.

Wu, H., Bailey, C., Rasoulinejad, P., Li, S., 2017. Automatic Landmark Estimation for Adolescent Idiopathic Scoliosis Assessment Using BoostNet. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. Springer International Publishing, pp. 127–135. doi:10.1007/978-3-319-66182-7_15.

Wu, Y., Yang, W., Lu, L., Lu, Z., Zhong, L., Huang, M., Feng, Y., Feng, Q., Chen, W., 2016. Prediction of CT substitutes from MR images based on local diffeomorphic mapping for brain PET attenuation correction. J. Nucl. Med. 57 (10), 1635–1641. doi:10.2967/jnumed.115.163121.

Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J., Li, S., 2017. Full Quantification of Left Ventricle via Deep Multitask Learning Network Respecting Intra- and Inter-task Relatedness. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 276–284. doi:10.1007/978-3-319-66179-7_32.

Yang, W., Zhong, L., Chen, Y., Lin, L., Lu, Z., Liu, S., Wu, Y., Feng, Q., Chen, W., 2018. Predicting ct image from MRI data through feature matching with learned nonlinear local descriptors. IEEE Trans. Med. Imaging 37 (4), 977–987. doi:10.1109/TMI.2018.2790962.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128.

Zhen, X., Wang, Z., Islam, A., Bhaduri, M., Chan, I., Li, S., 2014. Direct Estimation of Cardiac Bi-ventricular Volumes with Regression Forests. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. Springer International Publishing, pp. 586–593. doi:10.1007/978-3-319-10470-6_73.

Zhen, X., Wang, Z., Islam, A., Bhaduri, M., Chan, I., Li, S., 2016. Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Med. Image Anal. 30, 120–129. doi:10.1016/j.media.2015.07.003.

Zhen, X., Wang, Z., Yu, M., Li, S., 2015. Supervised descriptor learning for multi-output regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1211–1218.

Zhen, X., Zhang, H., Islam, A., Bhaduri, M., Chan, I., Li, S., 2017. Direct and simultaneous estimation of cardiac four chamber volumes by multioutput sparse regression. Med. Image Anal. 36, 184–196. doi:10.1016/j.media.2016.11.008.

Zhong, L., Lin, L., Lu, Z., Wu, Y., Lu, Z., Huang, M., Yang, W., Feng, Q., 2016. Predict ct image from MRI data using knn-regression with learned local descriptors. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 743–746. doi:10.1109/ISBI.2016.7493373.