Quantifying Axial Spine Images Using Object-specific Bi-path Network

Liyan Lin, Xi Tao, Wei Yang, Shumao Pang, Zhihai Su, Hai Lu, Shuo Li, Qianjin Feng* and Bo Chen

Abstract—Automatic estimation of indices from medical images is the main goal of computer-aided quantification (CADq), which speeds up diagnosis and lightens the workload of radiologists. Deep learning technique is a good choice for implementing CADq. Usually, to acquire high-accuracy quantification, specific network architecture needs to be designed for a given CADq task. In this study, considering that the target organs are the intervertebral disc and the dural sac, we propose an object-specific bi-path network (OSBP-Net) for axial spine image quantification. Each path of the OSBP-Net comprises a shallow feature extraction layer (SFE) and a deep feature extraction sub-network (DFE). The SFEs use different convolution strides because the two target organs have different anatomical sizes. The DFEs use average pooling for downsampling based on the observation that the target organs have lower intensity than the background. In addition, an inter-index dissimilarity constraint is proposed and applied to the output of the SFEs, taking into account that the activated regions in the feature maps of two paths should be different theoretically. An inter-index correlation regularization is introduced and applied to the output of the DFEs based on the observation that the diameter and area of the same object express an approximately linear relation. The prediction results of OSBP-Net are compared to several state-of-the-art machine learning-based CADq methods. The comparison reveals that the proposed methods precede other competing methods extensively, indicating its great potential for spine CADq.

Index Terms—Spine quantification, Bi-path network, Dissimilarity constraint, Inter-index correlation

I. INTRODUCTION

Over the last 30 years, medical imaging has greatly changed health care, enabling earlier detection of disease for patients, easier surveillance of disease for physicians, and greater efficiency in health care for hospitals. However, medical images produce large amounts of information, and physicians need to thoroughly analyze and evaluate the information in a short time. Heavy and urgent works promote the development of computer-aided diagnosis (CAD) technique. As a branch of CAD, computer-aided quantification (CADq) predicts indices of target organs, lesions, or tumors from medical images in real time [1], [2]. For example, axial spine CADq measures the intervertebral disc mid-sagittal diameter (IDMD), intervertebral disc cross-sectional area (IDCA), dural sac mid-sagittal diameter (DSMD), and dural sac cross-sectional area (DSCA) from axial spine images, as shown in Fig. 1(a).

Quantification of axial spine provides reliable means of disease diagnosis and assessment, such as spinal stenosis, degenerative disorders and lower back pain. Specifically, DSMD is measured for analyzing the mechanism of lower back-related leg pain [3]. DSCA is the gradation basis [4], [5] and the response criterion [6], [7] of spinal stenosis. DSMD also assists in determining the therapeutic effect of axial loading on a patient with degenerative disorders [6], the decompression effect of surgery on patients with spinal stenosis [8]–[10], or prognostic effect of a novel minimally invasive technique on patients with thoracic myelopathy [11]. IDMD is the diagnostic basis [12], and IDCA is the risk assessment [13] of lower back pain. IDMD and IDCA also help provide crucial information for surgical planning, artificial disc designing, and

Fig. 1: (a) Four indices of two target organs in axial spine MRI image. DSMD - dural sac mid-sagittal diameter; DSCA - dural sac cross-sectional area; IDMD - intervertebral disc mid-sagittal diameter; IDCA - intervertebral disc cross-sectional area. The multi-output regression problem makes the estimation of spine indices a challenging task. (b) Approximately linear relations exist between DSMD and DSCA and IDMD and IDCA. (c) Most part of the target organs have lower intensity than the background region. (d) High and low intensity disturbances in the background increase the difficulty of accurate estimation.
lower back biomechanical modeling [14].

In clinics, to acquire the four axial spine indices mentioned above, physicians need to trace the periphery or calculate the distance between the mid-sagittal extends of the intervertebral disc and vertebral disc areas in axial spine image [14]. Owing to the heavy manual procedure and the varying experiences among physicians, such an approach can result in unstable or even inaccurate measurement. To overcome this shortcoming and improve the stability and objectivity of prediction, researchers have been developing automatic methods to estimate spine indices based on machine learning techniques [15]–[18]. However, these methods have been designed for specific indices such as spine cobb angle and vertebral body height, which are anatomically different from the axial spine indices. Therefore, these methods require much improvement when they are applied for axial spine CADq. Moreover, deep learning obtains higher accuracy than conventional machine learning techniques in estimating indices [17], [19].

In this study, we propose an object-specific bi-path network (OSBP-Net) for axial spine CADq, which comprises two paths for two target organs, namely, the intervertebral disc and the dural sac. Each path contains a shallow feature extraction layer (SFE) and a deep feature extraction sub-network (DFE). The SFE in the intervertebral disc path uses a large convolution stride, whereas the SFE in the dural sac path uses a small one. The DFEs use average pooling instead of max pooling because the latter often ignores the object with low intensity. In addition, an inter-path dissimilarity constraint (IPDC) is proposed and applied to the output of the SFEs. The IPDC eliminates background and object disturbances by excluding the overlapped regions of features maps from two paths. An inter-index correlation regularization (IICR) is proposed and applied to the output of the DFEs to exploit the correlations between indices. The designs of OSBP-Net, IPDC, and IICR fully consider the intrinsic anatomical properties and gray value distributions of the axial spine images and the inherent correlation between different types of indices. As a consequence, the proposed method obtains quantities that are extensively superior to many other machine learning methods for axial spine CADq, as our experimental results demonstrated.

The contributions of this work are summarized as follows:
(1) The OSBP-Net comprising two paths for two target organs is proposed to estimate indices from axial spine images. The SFE in each path uses different convolution strides based on the consideration that the two target organs have different anatomical sizes. The DFEs use average pooling layers for downsampling based on the observation that the target organs have lower intensity than the background.
(2) The IPDC is proposed to constrain the feature extraction of the OSBP-Net. It is based on the consideration that the activated regions in the features maps of two paths should be different theoretically.
(3) The IICR is introduced to improve the predictions of the OSBP-Net. It is based on the observation that the diameter and area of the same object express an approximately linear relation.

The remainder of this paper is structured as follows. Section II reviews works related to estimation of spine indices and bi-path networks. Section III describes the architecture of OSBP-Net and the design of the SFE and DFE. Section IV gives the network loss function and the details of IPDC and IICR. Section V presents the experimental settings and the results. Finally, section VI draws conclusions.

II. RELATED WORKS

A. Estimation of Spine Indices

Manual measurement. In early years, some researchers manually measured spine indices in vitro [20] to evaluate the relationship between tissue excision and morphological change of intervertebral disc, and this method must be performed invasively. To date, the way of applying manual measurement has changed from measuring the vertebra in vitro to calculating the pixel numbers in regions of interests (ROIs) on spine images, and this change is promising for noninvasive spinal disease grading [4], [14]. However, regardless whether using MRI or CT image, usually tens or even hundreds of images should be calculated for a single patient. The manual measurement is very time consuming and laborious, reducing its availability for clinical or scientific practice.

Semi-automatic measurement. Some researchers reported using software to segment discs first, and then manually measure desired indices in the segment areas, resulting in semi-automatic methods [21], [22]. Despite saving some time and human resources, these methods still involve manual procedures. Moreover, the segmentation error influences the measurement results directly.

Automatic estimation. Recently, machine learning-based methods have emerged for automatically estimating indices from spine images. For example, Sun et al. [15] extracted histogram of directional gradient as feature descriptor from spinal X-ray image, and they constructed a support vector regression model to directly estimate cobb angle. Wu et al. [16] constructed a CNN model with BoostLayer to directly estimate landmarks and then mapped them to cobb angle. Pang et al. [17] incorporated the amplifier units into a cascaded CNN model to estimate the vertebra and disc height from sagittal spine MRI images, where the amplifier units selectively reuse the features between adjacent layers by a gate operation in skip connection. Wang et al. [18] constructed a CNN model with X-module to directly estimate cobb angle. The X-module exchanges feature maps from the other path with different views of images, which allows the network to learn some combined features from image of the other view. In summary, these models were specifically designed to predict indices from coronal X-ray or sagittal MRI images, but axial spine images have different anatomical structures. Therefore, their performance for axial spine CADq can be compromised. A dense enhancing network (DE-Net) has been very recently proposed to estimate axial spine indices and obtained relatively promising results [19]. However, the DE-Net is designed for general organ quantification, which does not exploit the intrinsic properties of axial spine estimation and therefore results in suboptimal predictions.
B. Bi-path Networks

The bi-path architecture is widely used in vision tasks, such as classification [23]–[25], segmentation [26], [27] and recognition [28]–[30], super-resolution [31], [32], quality evaluation [33], and sequence-to-sequence learning [34]. According to their applications, existing bi-path networks can be roughly classified into the following categories: 1) The siamese network is mainly applied by mapping different inputs to another space for object matching in retrieval tasks [28]–[30]. 2) The teacher-student network distills knowledge from a complex teacher network to a lightweight student network, which usually requires a large datasets for training [24], [25], [27]. 3) The dual branch-dual feature network can integrate the advantage the architecture of two branches for extracting different features [26], [31], [32], [34]. 4) The two-task network comprises two branches for two tasks where interactions are assigned to achieve mutual promotion between branches for feature extraction [18], [35].

These studies demonstrate that the bi-path architecture can be an optimal backbone when training a neural network to deal with two objects or sub-tasks. In our study, the indices to be estimated are related to the intervertebral disc and the dural sac in axial spine MRI image. Inspired by the success of bi-path architecture, we consider the intervertebral disc and the dural sac as two target objects. Following this idea, we construct the model for axial spine quantification as a bi-path network, with each path responsible for one object.

III. OBJECT-SPECIFIC BI-PATH NETWORK

A. Network Architecture

The architecture of the OSBP-Net is illustrated in Fig. 2(a). Splitting the network into two paths enables each path to focus on a single target (the intervertebral disc or the dural sac), therefore improving the network ability on discriminative feature embedding. Fig. 3 visualizes feature maps of two images from the 1st, 5th, 9th and 17th convolutional layers of two trained networks: a single path CNN and the proposed OSBP-Net. Evidently, the intervertebral disc and dural sac areas are more distinct in the feature maps from OSBP-Net compared with those from the single path CNN, thus demonstrating the superior ability of OSBP-Net on discriminative feature learning for axial spine CADq.

The two paths in OSBP-Net employ the same architecture. The only difference is that the convolution stride is 2 for SFE_{ID} and 1 for SFE_{DS}, which assists in extracting hierarchical features for the objects with different sizes. In forward propagation, the axial spine images are fed into the SFEs to perform shallow feature extraction. The outputs of the SFEs are then connected to the DFEs to extract deep features and each DFE produces a $2 \times 1$ prediction. In addition, the IPDC constrains the shallow feature maps generated by SFEs in the training phase, and the IICR regularizes the last fully connected layer of each DFE.
The parameters of OSBP-Net are listed in Table I. In intervertebral disk path, the first convolutional layer with a kernel size of 5 × 5 and stride of 2 provides large receptive field for the whole information extraction of the large target. In dural sac path, the first convolutional layer with a kernel size of 5 × 5 and stride of 1 ensures detail information extraction of the small target. At each scale, all convolutional layers in the convolution blocks with a kernel size of 3 × 3 correspond to a composite layer of “ReLU+BN+Conv” except for the first one. The last convolutional layer with a kernel size of 1 × 1 and stride of 1 linearly combines the feature maps for information integration. All average pooling layers with a size of 2 × 2 and stride of 2 are designed to extract translation-invariant feature except for the last one. The last average pooling layers with a size of 3 × 3/6 × 6 and stride of 3/6 are designed to reduce the dimensionality of feature maps, which are reshaped to a vector with a size of 4096 × 1 for regression. The first fully connected layer generates 2 × 1 coarse predictions for each path, which are then refined by the second fully connected layer with the regularization of IICR.

### B. Shallow Feature Extraction Layers

As illustrated in Fig. 2(b), SFE is a composite layer of convolution, batch normalization, and nonlinear activation. The convolution stride of the first layer significantly impacts the performance of the network, especially for small objects [36]. Anatomically, sizes of the intervertebral disc and the dural sac in the axial spine images are different. Such observation motivates us to use different strides in SFEs of different paths.

Specifically, the convolution stride is set as 2 for SFE of the intervertebral disc path and 1 for that of the other path.

For SFE$_{ID}$, as the target organ is relatively large, a large convolution stride is needed to enlarge the receptive field to extract its whole information. In the training phase, we adopt different convolution strides larger than 1 for SFE$_{ID}$ and find that a stride of 2 obtains the best prediction. For axial spine images, the dural sac area is approximately three times smaller than the intervertebral disc area. To ensure the SFE$_{DS}$ acquires more local detail information, a smaller convolution stride is required, which is set as 1 in this study.

Such strategy enables the SFEs to extract hierarchical features from spine axial images and therefore helps improve the prediction accuracy for each path of the OSBP-Net.

---

**Table I: Detailed parameters of the OSBP-Net.** Abbreviations: C = convolutional layer, CB = convolution block, AP = average pooling layer, FC = fully connected layer. The parameters of convolutional layers are presented as kernel size, stride, and number of filters. The parameters of average pooling layers are presented as pooling size, and stride.

<table>
<thead>
<tr>
<th>Layers</th>
<th>Layer parameter</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ID path</td>
<td>DS path</td>
</tr>
<tr>
<td>C0</td>
<td>5 × 5, 2, 8</td>
<td>5 × 5, 1, 8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB0</td>
<td>3 × 3, 1, 8</td>
<td>84 × 84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP0</td>
<td>2 × 2, 2</td>
<td>42 × 42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB1</td>
<td>3 × 3, 1, 16</td>
<td>42 × 42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP1</td>
<td>2 × 2, 2</td>
<td>21 × 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB2</td>
<td>3 × 3, 1, 128</td>
<td>21 × 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP2</td>
<td>2 × 2, 2</td>
<td>11 × 11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB3</td>
<td>3 × 3, 1, 512</td>
<td>11 × 11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP3</td>
<td>2 × 2, 2</td>
<td>6 × 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>1 × 1, 1, 1024</td>
<td>6 × 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP4</td>
<td>3 × 3, 6 × 6, 6</td>
<td>2 × 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reshaping</td>
<td>/</td>
<td>4096 × 1</td>
</tr>
<tr>
<td>FC0</td>
<td>/</td>
<td>2 × 1</td>
</tr>
<tr>
<td>FC1</td>
<td>/</td>
<td>2 × 1</td>
</tr>
</tbody>
</table>
C. Deep Feature Extraction Sub-networks

Fig. 4 illustrates the architecture of the DFE, which is a sequentially cascaded regression network. DFE comprises of 17 convolutional layers, 13 batch normalization layers, 12 ReLU activation layers, 5 average pooling layers, 1 feature reshaping layer, and 2 fully connected layers.

As shown in Fig. 1(c), in axial spine MRI images, the intervertebral disc area is darker and the dural sac area is partially darker compared with the background region. Based on this observation, instead of the commonly used max pooling for classification and regression tasks [17], [37]–[40], we adopt average pooling in the DFEs.

Note that max pooling is a very discrete operation by extracting the outstanding feature in pooling neighborhood, which may erase details from dark object [41]–[43], as illustrated in Fig. 2(d). Average pooling is a smooth operation compared with max pooling as it preserves information from the whole gray range after downsampling. The capacity of average pooling on encouraging network to identify all discriminative edges, which has been demonstrated to be effective in discriminative localization [44], is expressed as follows: The use of average pooling encourages the network to identify the complete extent of the object. The basic intuition behind this is that the loss for average pooling benefits when the network identifies all discriminative regions of an object as compared to max pooling.

Therefore, for the target organs that contain smaller gray value than the background in axial spine CADq, the average pooling strategy is more effective than max pooling. Fig. 5 shows exemplary feature maps of the 1st, 5th and 9th convolutional layers from two paths of the OSBP-Net with max or average pooling. The target organ regions are more distinct when the average pooling strategy is adopted, demonstrating its effectiveness for axial spine quantification.

IV. NETWORK LOSS WITH IPDC AND IICR

A. Network Loss

The loss function of OSBP-Net is formulated as

$$\mathcal{L} = \mathcal{L}_p + \alpha \mathcal{L}_{IPDC} + \beta \mathcal{L}_{IICR},$$

(1)

where $\mathcal{L}$, $\mathcal{L}_p$, $\mathcal{L}_{IPDC}$, and $\mathcal{L}_{IICR}$ denote the total network loss, primary loss, IPDC loss, and IICR loss, respectively; and $\alpha$, $\beta$ are hyper-parameters that control the regularization strength. The primary loss is formulated as

$$\mathcal{L}_p = \frac{1}{MN} \sum_{i=1}^{M} \left\| \hat{y}_i - y_i \right\|_1 + \frac{\gamma_{\theta}}{2} \sum \| \Theta \|_2,$$

(2)

where $\hat{y}_i$ and $y_i$ are the predicted and label indices vectors; $N$ is the number of indices to be estimated for each case ($N = 4$ for axial spine indices estimation); $M$ is the number of samples in a training batch; $\Theta$ denotes the set of network parameters; and $\gamma_{\theta}$ is a user-defined regularization parameter. The first term of the primary loss is a $\ell_1$ norm of the prediction errors, which forces the network output to be as close as possible to the true indices. The second term is a $\ell_2$ norm regularization for the network parameters, which is used for alleviating the over-fitting problem.
B. Inter-path Dissimilarity Constraint

Under the constraint of IPDC, feature maps in the homologous layers of two paths are encouraged to be spatially different, which guides the network to up-weight contribution of the feature from target organ region and down-weight that from background region.

The theory behind the idea of IPDC is that, on the one hand, for the object region, pixels belong to one object have no intersection with those belong to the other one in the image. Ideally, features in one path only highlight pixels associated with the corresponding object. Therefore, we assume that the intersection of object regions in feature maps of the homologous layers from two paths should be as low as possible. On the other hand, for the background region, feature maps in both paths contain the similar background information. In theory, features in both paths contain no background information. As such, we assume that the intersection of background region in feature maps of the homologous layers from two paths should be as low as possible.

Considering that high-resolution feature maps generated by the first layer contain the most low-level spatial information, we apply the IPDC to the feature maps outputted by SFEs. By doing so, the disturbances in background region will be suppressed, and the object regions will be highlighted.

Specifically, for the OSBP-Net with an intervertebral disc path \( F_{ID}(I) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{2 \times 1 \times 1 \times D} \), and a dural sac path \( F_{DS}(I) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{2 \times 1 \times 1 \times DS} \), we calculate the cosine similarity between the feature maps generated by SFEs as follows:

\[
\mathcal{L}_{IPDC}^I = \frac{\mathcal{F}^{C0}_{ID}(I) \cdot \mathcal{F}^{C0}_{DS}(I)}{\|\mathcal{F}^{C0}_{ID}(I)\|_2 \|\mathcal{F}^{C0}_{DS}(I)\|_2},
\]

To avoid the influence of intensity difference between \( \mathcal{F}^{C0}_{ID} \) and \( \mathcal{F}^{C0}_{DS} \), the feature maps are normalized before matching:

\[
\mathcal{N}^{C0}(I) = \frac{\mathcal{F}^{C0}(I)}{\|\mathcal{F}^{C0}(I)\|_p},
\]

where \( \|\cdot\|_p \) denotes \( p \) norm of the feature map vector. Here, we use \( \ell_1 \) norm, which empirically works well. Then, the IPDC loss function can be formulated as

\[
\mathcal{L}_{IPDC} = \frac{\mathcal{N}^{C0}_{ID}(I) \cdot \mathcal{N}^{C0}_{DS}(I)}{\|\mathcal{N}^{C0}_{ID}(I)\|_2 \|\mathcal{N}^{C0}_{DS}(I)\|_2}.
\]

Through minimizing (5), the intersection of feature maps between two paths can be reduced, which improves the attention of each path on its target organ region. The mechanism of IPDC is illustrated in Fig. 2(b).

Seeing that the IPDC has the ability to improve the attention of OSBP-Net, we intend to compare the performance of IPDC with popular attention-based methods, including the non-local block (NLB) [45] and the spatial attention module (SAM) [46].

Fig. 6 visualizes the shallow feature maps of the ID and DS paths obtained by the naive OSBP-Net and OSBP-Net with NLB, SAM, or IPDC. Incorporating any of these additives into the OSBP-Net can suppress the disturbances and highlight the target areas. Among them, the OSBP-Net with the proposed IPDC generates the best visualization.

C. Inter-index Correlation Regularization

As illustrated in Fig. 1(b), the distance and area of one object express an approximately linear relation. To leverage the latent correlation between the indices of the same object for better regression performance, a structure matrix is constructed to model index correlations.

Specifically, we denote the last fully connected layer of the DFEs as \( \mathcal{F}^{FC1}_{ID}(\cdot) : \mathbb{R}^{2 \times 1} \rightarrow \mathbb{R}^{2 \times 1 \times ID} \) and \( \mathcal{F}^{FC1}_{DS}(\cdot) : \mathbb{R}^{2 \times 1} \rightarrow \mathbb{R}^{2 \times 1 \times DS} \), respectively. Elements of structure matrices \( S_{ID} \) and \( S_{DS} \) are the weights of the fully connected layers \( \mathcal{F}^{FC1}_{ID}(\cdot) \) and \( \mathcal{F}^{FC1}_{DS}(\cdot) \), which can be represented as \( \{S_{i,j}\}_{ID}^{FC1} \), \( \{S_{i,j}\}_{DS}^{FC1} \). where \( w_{i,j} \) is the weight for \( i \)th output and \( j \)th input. The latent correlation between the distance and area of the same object is explicitly encoded via \( \ell_\infty \) norm based learning method, which is formulated as

\[
\mathcal{L}_{IICR} = \|S_{ID}\|_\infty + \|S_{DS}\|_\infty, \tag{6}
\]

\[
\|S\|_\infty = \max_{1 \leq i \leq r} \sum_{j=1}^{c} |S_{i,j}|, \tag{7}
\]

where \( S_{ID} \) and \( S_{DS} \) are the structure matrices for the intervertebral disc path and the dural sac path, respectively; \( r \) and \( c \) are the total number of rows and columns, respectively; and \( S_{i,j} \) are the \( i \)th row and the \( j \)th column of matrix \( S \).

Under the regularization of IICR, the structure matrices provide predicted indices with similar sparse parameter through \( \ell_\infty \) minimization, which assists in acquiring common features for the final prediction. Thus, the performance of the OSBP-Net can be optimized by leveraging knowledge across correlated indices that share common features. Essentially, IICR is an indirect modeling of image feature, and we apply infinity norm for modeling the relationship between the diameter and area of one organ. Other matrix norms are applicable, too.
V. Experiments and Results

A. Experimental Settings

1) Dataset: The dataset consists of 895 axial spine MRI images from 143 patients collected from a Philips Medical Systems-Ingenuity MRI machine with the following specific parameters: echo time 120 ms, repetition time 3228.15 ms, flip angle 90°, and T2 weight TSE DRIVE HR series.

The patients include 72 females and 71 males with age ranging from 37 to 69. For each patient, 6 ± 2 images were manually selected and measured by two physicians with more than eight years of experience using ITK-SNAP tool [47]. The average of the measurements obtained by two physicians was used as the ground truth. The grid of the images is 336×336 with the pixel size of 0.53×0.53 mm². Before training, all images undergo ROI cropping using a bounding box with the fixed top-left and bottom-right pixel coordinate of (84, 217) and (251, 50) and then further cropped as a squared region around disc and dural sac with a size of 168×168. This study is approved by the Institutional Review Board of the Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China on Jan. 9th, 2020 (protocol number: K05-1).

2) Implementation: In our experiments, all networks were implemented by Tensorflow on a Linux server with a Tesla P100-SXM2 GPU. The SGD solver was adopted to propagate the errors for updating the network parameters. For the proposed OSBP-Net, the training parameters were set as weight decay 0.001, Nesterov momentum 0.9, number of epochs 500, and batch size 8. The learning rate was set as 0.01 initially and reduced by 10 folds at epochs 150 and 400. All experiments were done on 2D slices.

3) Evaluation: To comprehensively evaluate all competing models, 5-fold cross validation was employed. The dataset was divided into 5 groups by randomly dividing patients. Four groups were used as the training set and the last group was used as the validation set in turn. The mean absolute error (MAE) between the predicted and true indices was computed as:

$$\text{MAE}(\hat{y}, y) = \frac{1}{K} \sum_{k=1}^{K} |\hat{y}_k - y_k|,$$  \hspace{1cm} (8)

where $K$ is the number of indices to be computed in a set.

B. Parameter Selection

1) Kernel size: In the proposed OSBP-Net, the IPDC constrains shallow features generated by SFEs, making the SFEs unique compared with layers in the DFEs. In our study, we tested the performance of OSBP-Net using SFEs with kernel size of 3×3, 5×5, and 7×7. The results are presented in Fig. 7. For all indices, OSBP-Net using SFEs with a kernel size of 5×5 constantly obtains the smallest errors. Therefore, the kernel size of SFEs was set as 5×5 for the rest of the experiments. The kernel size of convolution layers in DFEs was empirically set as 3×3, which is same as many classification or regression networks [17], [37]–[40].

2) Depth of DFEs: As the depth of DFEs indicates the depth of the OSBP-Net, we tested the performance of OSBP-Net using SFEs with different depths. Fig. 8 shows the MAEs obtained by the OSBP-Net with DFEs of 2, 3, and 4 scales (each scale has 4 convolutional layers) and deeper DFEs with one or two more convolutional layers following 4-scale DFE (4+1 Conv, 4+2 Conv). The figure reveals that OSBP-Net with 4-scale DFE obtains the smallest MAE on average. Therefore, for all other experiments, both the number of convolutional layers in each scale and the number of scales were set as 4 for the DFEs.

3) Regularization parameters: To select optimal regularization parameters for the network loss function, we trained the OSBP-Net multiple times with different regularization parameters. The results are presented in Fig. 9, wherein 0.001 for $\gamma_1$, 0.0005 for $\alpha$, and 0.01 for $\beta$ obtain the smallest test error. These values of regularization parameters were also adopted for other experiments.

C. Ablation Studies

1) Effectiveness of bi-path architecture: To verify the effectiveness of splitting the network into two paths, we first compare the predictions of single-path CNN and the OSBP-Net. As can be seen from the first and second rows of Table II, the OSBP-Net acquires lower MAE than single-path CNN. In addition, we compare the visualizations of exemplary feature maps obtained from the single-path CNN and the OSBP-Net in Fig. 3, where the OSBP-Net is able to extract more discriminative features from the axial spine images.

2) Effectiveness of using different convolution strides in SFEs: Comparing the second and third rows of Table II, the prediction errors can be reduced by using SFEs with different convolution strides. The numbers in the second row are produced by OSBP-Net using SFEs with the same stride of 2 and a kernel size of 7×7. The numbers in the third row are produced by OSBP-Net using SFE with a stride of 2 and SFE_DS with a stride of 1 as well as a kernel size of 5×5 in both paths. Moreover, different combinations of convolution stride for SFEs are tested, and the results are compared in Fig. 10. A stride of 1 is optimal for SFE_DS and a stride of 2 is optimal for SFE_ID, which further demonstrates the benefit of using different convolution strides in SFEs.

3) Effectiveness of using average pooling strategy in DFEs: Comparing the third and fourth rows of Table II, OSBP-Net using DFEs with average pooling strategy can result in smaller MAEs. The numbers in the third row are produced by OSBP-Net using DFEs with max pooling strategy. The feature...
TABLE II: MAE results of ablation studies. SFE\textsubscript{DCS}: using SFEs with different convolution stride. DFE\textsubscript{APS}: using DFEs with average pooling strategy. The MAEs become smaller and smaller with the proposed strategies being integrated into the OSBP-Net gradually.

<table>
<thead>
<tr>
<th>Bi-path</th>
<th>SFE\textsubscript{DCS}</th>
<th>DFE\textsubscript{APS}</th>
<th>IPDC</th>
<th>IICR</th>
<th>DSAD (mm)</th>
<th>DSCA (mm\textsuperscript{2})</th>
<th>IDAD (mm)</th>
<th>IDCA (mm\textsuperscript{2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>1.061 ± 0.119</td>
<td>2.327 ± 0.166</td>
<td>1.381 ± 0.085</td>
<td>3.342 ± 0.287</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>1.055 ± 0.096</td>
<td>2.377 ± 0.164</td>
<td>1.376 ± 0.166</td>
<td>3.268 ± 0.266</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>1.005 ± 0.103</td>
<td>2.336 ± 0.194</td>
<td>1.315 ± 0.069</td>
<td>3.163 ± 0.264</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.935 ± 0.093</td>
<td>2.274 ± 0.169</td>
<td>1.277 ± 0.057</td>
<td>3.143 ± 0.287</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.922 ± 0.075</td>
<td>2.229 ± 0.187</td>
<td>1.276 ± 0.084</td>
<td>2.977 ± 0.206</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.868 ± 0.063</td>
<td>2.182 ± 0.152</td>
<td>1.269 ± 0.081</td>
<td>2.977 ± 0.265</td>
</tr>
</tbody>
</table>

Fig. 8: MAEs obtained by the OSBP-Net with DFEs of 2, 3, and 4 scales (each scale has four convolutional layers) and deeper DFEs with one or two more convolutional layers following 4-scale DFE (4+1 Conv, 4+2 Conv).

maps visualized in Fig. 5 also reveal that the average pooling strategy is a better choice than max pooling for extracting distinct edges of the target organs from axial spine MRI images.

4) Effectiveness of IPDC: Comparing the fourth and fifth rows of Table II, OSBP-Net with IPDC reduces the prediction errors for all indices in comparison with OSBP-Net without IPDC. Given this, we compare the results of the naive OSBP-Net and OSBP-Net with NLB [45], SAM [46], or IPDC. Table III compares the prediction errors obtained by OSBP-Net with NLB, SAM, or IPDC. The results of OSBP-Net with NLB are not better than the naive OSBP-Net, indicating the ineffectiveness of NLB for the given task. The OSBP-Net with SAM obtains lower errors than that with the proposed IPDC on DSAD and IDAD prediction. The OSBP-Net with IPDC acquires the best predictions on average. Moreover, the feature maps shown in Fig. 6 indicates that the OSBP-Net with the proposed IPDC generates the best visualization among the four configurations compared. These numerical and visual results demonstrate that the IPDC is especially effective for the task we intend to accomplish.

5) Effectiveness of IICR: Comparison of the fifth and sixth rows of Table II tells that training OSBP-Net with IICR could improve prediction accuracy for three indices except for the IDCA. Furthermore, Fig. 11 shows the curves of average training loss and test error versus the training epoch of all cross experiments with and without the IICR. The curves of training loss are very close for both models with and without IICR. In the testing stage, the model with IICR obtains test error far below the one without IICR. This observation demonstrates that the IICR can alleviate the over-fitting problem and improve the prediction accuracy effectively.

The MAEs listed in Table II also reveals that, by gradually integrating the proposed strategies into the network, the prediction errors become increasingly smaller.

D. Overall Performance

Fig. 13 compares the estimation and manual measurement for four indices of 100 randomly selected axial spine images. The estimated indices are very close to the manually measured ones. The prediction errors of our proposed OSBP-Net are 0.87 ± 0.06 mm, 2.18 ± 0.15 mm\textsuperscript{2}, 1.27 ± 0.08 mm, and 2.98 ± 0.27 mm\textsuperscript{2} for the DSMD, DSCA, IDMD, and IDCA, respectively. Note that these numbers are very small compared with the true values. Therefore, the axial spine indices can be accurately estimated by the proposed method, which is very helpful for analyzing the pathological and anatomical changes of spines. Furthermore, as these indices are anatomically correlated with each other, the variational patterns of the DSMD and DSCA are similar as well as the IDMD and IDCA. Fig. 13 reveals that the estimated indices also possess these similarities of variational patterns, indicating that the proposed
method can effectively learn the patterns between different indices and samples.

In addition, we compare the results of the OSBP-Net with results obtained by Multi-features+RF [48], HOG+AKRF [49], HOG+SSVR [15], the cascaded amplifier regression network (CARN) [17], and the dense enhancing network (DE-Net) [19]. The Multi-features+RF is designed for cardiac estimation with all indices being areas (biventricular volumes). The HOG+AKRF is designed for head post estimation. The HOG+SSVR is designed for spinal cobb angle estimation. The CARN is designed for sagittal spine indices estimation with all indices being distances (the vertebral body height and the intervertebral disc height). The DE-Net is one-path network designed for estimating indices from axial spine image. To ensure fair comparison, parameters of all competing methods are tuned to acquire the smallest MAE values. Fig. 12 shows the MAEs obtained by Multi-features+RF, HOG+AKRF, HOG+SSVR, CARN, DE-Net, and OSBP-Net with respect to the four axial spine indices. All deep learning-based methods gain results better than the conventional machine learning-based methods. Overall, the proposed OSBP-Net obtains the smallest errors for all indices among all competing methods, which demonstrates the superiority of OSBP-Net against other models.

VI. CONCLUSIONS

In this study, we performed the axial spine CADq task using an object-specific bi-path network with IPDC and IICR. The OSBP-Net, IPDC, and IICR were designed by fully exploiting the intrinsic anatomical properties and gray value distributions of the axial spine images and the inherent correlation between different types of quantitative indices. Comprehensive experiments were conducted to verify the effectiveness of all proposed strategies for axial spine quantification. Performance comparison with many other CADq models demonstrated the effectiveness of the proposed method on axial spine CADq. Future works mainly lies in performing spine CADq using 3D, multi-modality, or multi-view images to further improve the prediction accuracy.

ACKNOWLEDGEMENT

We performed all experiments in 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 by the Ontario Government and the Federal Economic Development Agency for Southern Ontario.

REFERENCES

Fig. 13: Estimations versus manual measurements for four indices of 100 randomly selected axial spine images.


