IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 10, 2021, accepted February 15, 2021, date of publication February 22, 2021, date of current version March 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061090

Learning-Based Coronal Spine Alignment Prediction Using Smartphone-Acquired Scoliosis Radiograph Images

TENG ZHANG[1], (Member, IEEE), YIFEI LI[2], JASON PUI YIN CHEUNG[1], SOCRATES DOKOS[3], (Member, IEEE), AND KWAN-YEE K. WONG[2], (Senior Member, IEEE)

[1]Department of Orthopaedics and Traumatology, The University of Hong Kong, Hong Kong
[2]Department of Computer Science, The University of Hong Kong, Hong Kong
[3]Graduate School of Biomedical Engineering, University of New South Wales, Sydney, NSW 2052, Australia

Corresponding author: Jason Pui Yin Cheung (cheungjp@hku.hk)

This work was supported in part by the Innovation and Technology Fund under Grant ITS/404/18, and in part by the AOSpine East Asia Fund under Grant AOSEA(R)2019-06.

ABSTRACT DICOM X-rays are not easily accessible for telemedicine, and existing learning-based automated Cobb angle (CA) predictions are not accurate on suboptimal X-ray images. To develop an automated CA prediction system irrespective of image quality, with no restrictions on curve patterns, 367 consecutive patients attending our scoliosis clinic were recruited and their coronal X-rays were re-captured using mobile phones. Five-fold cross-validation was conducted (each with 294 randomly selected images for training a neural network SpineHRNet to detect endplate landmarks and end-vertebrae, and the remaining 73 images for testing). The predicted heatmaps of vertebral landmarks were visualized to enhance interpretability of the SpineHRNet. Per-landmark Euclidean distance (L2) errors and recall of landmark detection were calculated to assess the accuracy of the predicted landmarks. Further computed CAs were quantitatively compared with spine-specialists measured ground truth (GT). The average L2 error and the recall of the detected endplates landmarks were 2.8 pixels and 0.99 respectively. The predicted CAs were all significantly correlated with GT ($p<0.01$). Compared with GT, the mean absolute error was 3.73-4.15° and standard deviation was 0.8-1.7° for the predicted CAs at different spinal regions. This is the first study on non-original X-rays to automatically and accurately predict endplate landmarks of the scoliotic spine and compute the CAs at different regions of the spine, irrespective of image qualities. SpineHRNet's applicability is evidenced by five-fold cross-validations, which may be used with telemedicine to facilitate fast and reliable auto-diagnosis and follow-up.

INDEX TERMS Automatic analysis, computer vision, HRNet, telemedicine, landmark detection, out of hospital consultation.

I. INTRODUCTION

Adolescent idiopathic scoliosis (AIS) is the most common pediatric spinal deformity [1], characterized by lateral curvature of the spine [2], [3] on coronal X-rays [3], [4]. If untreated, curve progression can reach 90% [5], [6]. Up to 38% of patients progress, despite following brace-wear-protocol [7], [8], thus careful follow-ups are critical. Cobb angles (CAs), which are measures of spine curvature in degrees, are the primary consideration for AIS diagnosis before appropriate treatment planning can be conducted [9]. To measure the CAs, the end vertebrae need to be identified, which are the most tilted vertebrae away from the horizontal apical vertebra. The CA is then measured by the angle formed by lines drawn at the superior and inferior endplates of the upper and lower end vertebrae respectively. Previous studies have demonstrated traditional image processing techniques [10]–[18] for feature extraction and CA calculation [16]. Due to the heterogeneous patterns of deformities (i.e., different curve locations and combinations) and X-rays having high variance (due to different equipment with different technicians), these methods have limited accuracy and are not applicable for direct clinical use.

Recent advances in artificial intelligence (AI) CA automation [19]–[22] can directly [23] or indirectly [21], [22] determine CAs from X-rays limited to a single curve, but cannot

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

handle heterogeneous patterns of curves. It is also difficult to guarantee the model has learnt the correct computation of CA without intermediate supervision [23].

Deep learning analysis of CAs utilizing convolutional neural networks (CNNs) indicated the possibility of automated spinal shape detection, but with unsatisfied accuracy despite the use of original high-resolution X-rays [24].

The recently proposed human pose estimation network High-Resolution Net (HRNet) [25] can improve the accuracy of landmark detection in natural images. Most existing landmark detection networks consist of several cascaded encoder-decoder submodules, which down-sample and up-sample feature maps sequentially. HRNet, on the contrary, can maintain high-resolution representations through the whole network. It gradually adds sub-branches with low-resolution representations in a parallel manner, and fuses multi-scale features in its final high-resolution representation. It may therefore have applications in medical imaging analysis when accurate key point landmark detections are required, as is the case for endplate landmark detection.

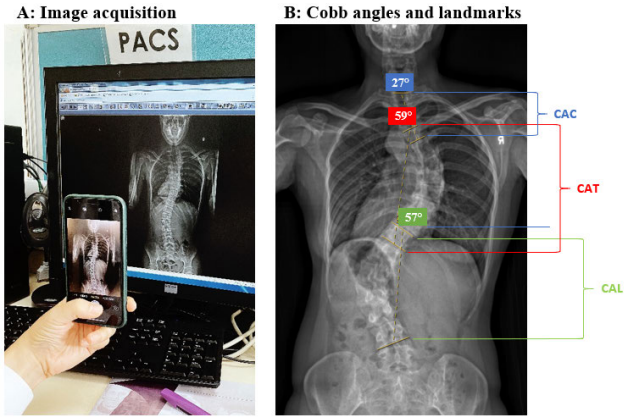
The stationary picture archiving and communication system (PACS) is conventionally used for viewing and manually assessing DICOM X-rays with built-in manual tools, which are not easily accessible or modifiable. However, to facilitate real-time or out of hospital follow-up, it is popular for spine specialists to take a photo of the X-ray with a smartphone for further communication with other clinicians and patient carers [26], [27]. An automatic tool for accurately detecting vertebrae landmarks on images of various quality can provide an easily accessible tool to evaluate deformities.

This study aims to provide reliable automated vertebral landmark detection, irrespective of image quality, thus, to potentially facilitate real-time diagnosis or out of hospital follow-up. The objectives are, 1) to establish a reliable deep learning-based method to accurately detect vertebral landmarks, including endplates and end vertebrae; 2) to eliminate previous restrictions of automatic coronal alignment on curve patterns or imaging quality by training the model using non-original X-rays of various image quality and different curve patterns; and 3) to examine the vertebral landmarks detection accuracy and CA computation accuracy by comparing with the specialists measurements.

II. MATERIAL AND METHODOLOGY

A. DATASET PREPARATION AND IMAGE PRE-PROCESSING

Images of X-rays from 367 consecutive AIS patients (80% female; age 10-18) who visited our clinic underwent screenshot of their X-rays by smartphones (Fig. 1A: including iPhone 8 and iPhone 8 Plus; Apple Inc.) from April to June of 2019. This study was ethically approved by the local institutional review board. Patients were excluded if they had psychological and/or systematic neural disorders that could influence the compliance of the study and/or patient mobility (e.g., prior cerebrovascular accident, Parkinson's disease, myopathy), congenital spinal deformities, previous



C: Deformity severity and clinical relevance of Cobb angles

CAs range	Severity and clinical management
0°-20°	Normal - Mild
	No specific interventions required, but for cases with curves between 10°-20°, follow-up may be required annually.
21°-40°	Moderate
	Nonoperative management would be suggested, with no invasive procedures, including brace wearing and scoliosis specific exercises.
>41°	Severe
	Surgeries maybe required to correct the spine deformity

FIGURE 1. Example of the image acquisition process, end vertebra and Cobb angles. The images were acquired by using smartphones and screenshots of the X-rays displayed on the PACS (A). Cobb angles (CAs) measured by the angles formed by lines drawn at the upper and lower endplates of the upper and lower end vertebrae (the most tilted vertebrae from the apical vertebra) respectively, with CAC, CAT, CAL representing CAs at different regions of the spine (B). The deformity severity of the spine is classified according to the CAs, with 0°-20° being normal to mild, 21°-40° being moderate and over 40° being severe (C). Different clinical interventions ranging from nonoperative management to surgeries, would be required according to different severities.

spinal operations, any trauma that could impair posture and mobility, and any oncological diseases. The technicians were instructed to take photos or screenshots of the displayed X-ray while maintaining the image plane parallel to the screen, excluding the patient's demographic information from the capturing field to anonymize the X-rays. All images were uploaded to our internal server via an in-house developed mobile application.

The image collection was followed by labelling 4 endpoints of the 2 endplates of each vertebral body, manually by spine specialists, using a self-developed Python-script, for key point placement on the images. The upper 6 cervical vertebrae were occluded by the skull, leaving 18 distinguishable vertebrae from the 7^{th} cervical vertebra to the 5^{th} lumbar vertebra (C7-L5), rising to $18 \times 4 = 72$ endplate landmarks. End vertebrae and CAs (Fig. 1B), manually assessed by spine specialists for deformity severity diagnosis and treatment planning (Fig. 1C), were considered as ground truth (GT). The inter-rater variation between the two specialists who labelled the images were tested on fifty images. To clarify the position of each curve, the CAs were triaged according to whether the curvature started in the cervico-thoracic region (CAC), the thoracic region (CAT) or the lumbar region (CAL).

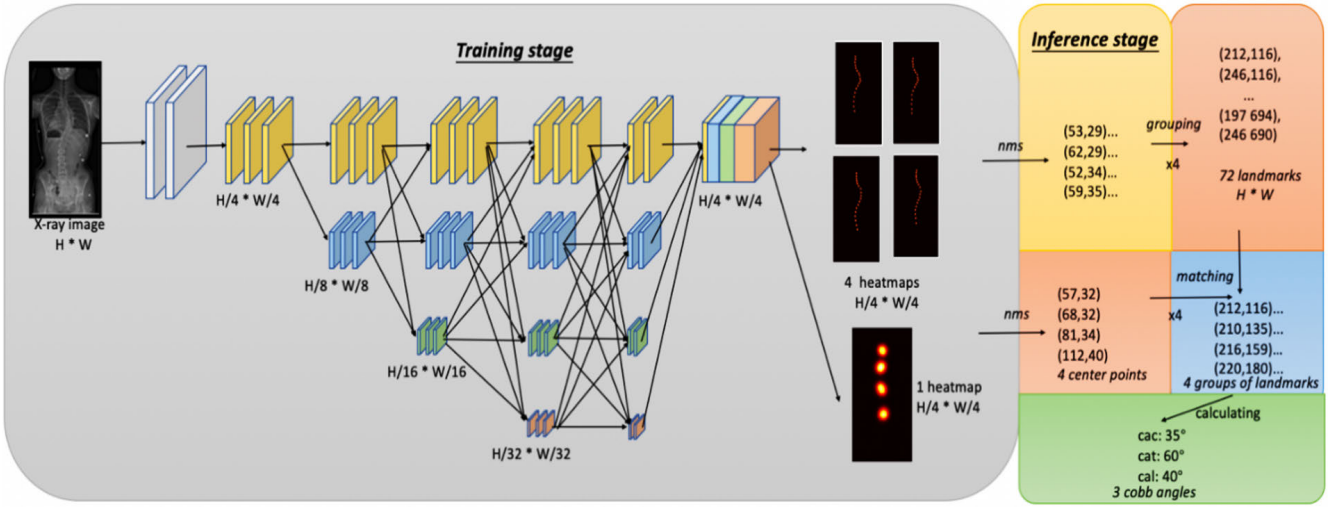


FIGURE 2. The overall pipeline of the automated Cobb angles using SpineHRNet. The left panel (grey) shows the network architecture, which predicts 4 heatmaps for the 4 endpoints of the vertebrae/endplates and 1 heatmap for the locations of the end vertebrae. The right panel illustrates the inference stage with different sub-stages. During inference, endpoint locations of the endplates were extracted by Non-Maximum-Suppression (NMS) and matched to different vertebrae, and the center point locations of the end vertebra were extracted by NMS from the end-vertebra-heatmap. Subsequently the end vertebra center point with the nearest 4 endplate landmarks were identified to calculate the CAs.

The image quality varied with different resolution, intensity and anatomical structures contained in the X-rays. The image height range was 654-892 pixels (mean=887.3; and median=892.0), whilst width range was 386-1384 pixels (mean=704.8; median=696.0), and the mean intensity range per image was 23.8-116.8, with most of images containing the whole spine, but a few also including the whole body (with the lower limbs). Due to the large variance in the collected dataset, pre-processing was performed to automatically remove the surrounding background pixels (i.e., without affecting the spine) created by the optical acquisition of the X-ray images displayed on the PACS. After the auto-cropping, the range, mean, and median of image height were 654-892, 887.3 and 892 pixels respectively; with the range, mean and median of image width being 318-893, 439.2 and 430 pixels respectively. To further unify the size of the input images, the re-captured X-ray images were automatically cropped and resized with zero padding to a fixed dimension of 896 $\times$ 448 pixels containing the whole spine.

B. DEEP LEARNING-BASED VERTEBRAL LANDMARKS DETECTION

Our new approach consisted of a two-stage detection design to identify the vertebral landmarks, including 1) the endplate landmarks and 2) the end vertebrae. Thus, the CAs were computed based on pre-detected endplate landmarks of the end vertebrae. With this two-stage formulation, it was not necessary to fix the number of CAs, as both the patterns of curves and the number of CAs could be inferred from the detected end vertebrae (Fig. 2). Importantly, these endplate landmarks were close to each other in the adjacent vertebrae. Therefore, it was essential to utilize high-resolution feature maps with sufficient low-level information, which is the advantage of HRNet [25]. Additionally, locating landmarks in the form of heatmaps has been shown to be more effective and accurate than directly regressing coordinates [28].

C. HEATMAP GENERATION AND SUPERVISION

To detect the endplate landmarks and end vertebrae, we utilized heatmap representation as our supervision target. The advantages of a heatmap include: 1) it can better capture ambiguities in landmark labelling since the pixels around the labelled landmarks are probable landmarks; 2) directly outputting coordinates is a highly non-linear mapping from the image to quantitative numbers; 3) the heatmap can be visualized, which serves as an interpretable guide for clinicians.

For the endplate landmark detections (Fig. 2), heatmaps were generated as a 2D Gaussian distribution centred at each of the ground-truth landmarks, where the pixel value indicated the probability of it being a landmark. Furthermore, the landmark estimation was not formulated as a single-spine landmark estimation, but multi-vertebra landmark estimation through a bottom-up approach [29]. If the single-spine landmark estimation approach were adopted to detect 72 landmarks, we would be generating 72 heatmaps, each corresponding to one landmark of one end of the endplate. However, we adopted the multi-vertebra landmark estimation approach, thus we generated 4 heatmaps, each corresponding to one endpoint of one endplate of one end vertebra. For example, the first heatmap corresponded to the left-upper landmarks and the last heatmap corresponded to the right-lower landmarks of all the endplates. As a result, the heatmaps were generated from a multi-peak 2D Gaussian distribution ($\sigma = 1$, where σ is the standard deviation of the Gaussian distribution: this value was selected due the necessity of key point detection of the endplate landmarks).

For end vertebrae detections (Fig. 2), we also adopted such multi-peak heatmaps as a supervision target, enabling the model to detect variable numbers of end vertebrae. Therefore, we only generated 1 heatmap for all the end vertebra, with each peak value indicating the center of one end vertebra, defined as the average of four endplate landmarks of this end vertebra, with the σ value set to 6 (decided empirically).

D. DATA AUGMENTATION AND TRAINING POLICIES

To further mimic the real-world situation and enhance the robustness of our network to handle images captured under different setups and quality, and to avoid overfitting our dataset, extensive data-augmentation during training was carried out. We did not use a fixed augmented dataset but conducted augmentation in the training procedure. With this augmentation policy, the size of our augmented dataset was unbounded. Specifically, the image was firstly read into memory, followed by a random flip (probability=0.5), random scale ([0.8, 1.2]), random rotation ([−5°, 5°]), random horizontal translation ([−75 pixels, 75 pixels]), random vertical translation ([−10 pixels, 10 pixels]), and a random contrast augmentation ([0.8, 1.2]). Furthermore, we conducted additional cropping or padding to ensure the size of images was fixed to 896 × 448 pixels, since the augmentation would change the image size. The generated heatmaps and GT landmarks were correspondingly transformed by the same augmentations as the input images, with the mini-batch size being 16.

For the heatmap generation described previously, the supervision process could be considered as a classification problem rather than a regression problem. We also tested empirically, and found that the binary cross entropy loss (BCE) (1), as shown at the bottom of the page, yielded more accurate predicted heatmaps than the regression loss in L1 (2), as shown at the bottom of the page, distance (the sum of distance error in x and y axes) and L2 (3), as shown at the bottom of the page, distance (the per-landmark Euclidean distance error).

where B denotes the mini-batch-size, i the sample index in each batch, (H, W) the sample shape, (x, y) the spatial coordinates, G the GT heatmaps, and O the predicted heatmaps. Although the network outputs a heatmap with resolution down-sampled by a factor of 4, all experiment results are based on the full resolution by multiplying 4 to the coordinates of the predicted landmarks.

We trained our model with the Adam optimizer [30] by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$ for both landmark estimations and end-vertebrae detection. For the endplate landmarks estimation, the base learning rate is set as $1e^{-2}$, dropping to $1e^{-3}$ and $1e^{-4}$ at the 30th and 50th epochs. Besides, the learning rate was gradually increased to $1e^{-2}$ from $1e^{-3}$ in the first 10 epochs (namely learning-rate warm-up). The training process was terminated at 100 epochs.

To detect the end vertebra, the experiment settings were similar, except that the initial learning rate was set to $2e^{-3}$; subsequently dropping to $2e^{-4}$ and $2e^{-5}$, respectively, at the 30^{th} and 50^{th} epochs. We initialized both networks from the ImageNet-pretrained checkpoint offered by HRNet model zoo. Similarly, the end vertebrae detection network was initialized with the trained endplate landmark estimation network. We implemented our models in the PyTorch framework, training these using 4 NVIDIA Titan Xp GPUs. The network can be downloaded from the following link (https://github.com/rovephoenix/automated-spine-analysis).

E. INFERENCE STAGE AND COBB ANGLE CALCULATION

At the inference stage, we firstly located every peak value from each heatmap using the Non-Maximum-Suppression (NMS) algorithm. For each end vertebra peak, 4 associated endplate landmarks were grouped, by aligning the closest 4 endplate landmarks with the center points of the predicted end vertebrae (Fig. 2). CAs were calculated via the detected top endplate of the predicted top end vertebra, and the bottom endplate of the predicted top end vertebra.

F. RELIABILITY ASSESSMENTS

Reliability assessments were conducted in a 5-fold cross-validation manner to ensure a comprehensive and reliable evaluation. The dataset was split into 5 exclusive folds (73 images/fold). 5 independent experiments were conducted, of which a randomly selected 4 folds were used for training and 1-fold for testing. For the accuracy of landmark detections, we evaluated the landmark retrieval rate and the per-landmark Euclidean distance (L2) errors. For CA predictions, the recall, precision, and F1-score were evaluated to measure the retrieval performance of our method. The absolute error between the GT and the SpineHRNet predicted results were evaluated. The prediction reliability was tested by regression analysis and Bland-Altman plots of the GT with the results pooled from the 5-fold cross-validation.

$$BCE = -\frac{\Sigma_i^B \Sigma_{x,y}^{H,W} \left[G_i[x,y] \times \log O_i[x,y] + (1 - G_i[x,y]) \times \log (1 - O_i[x,y]) \right]}{H \times W \times B} \tag{1}$$

$$L1 = \frac{\Sigma_i^B \Sigma_{x,y}^{H,W} \| G_i[x,y] - O_i[x,y] \|_1}{H \times W \times B} \tag{2}$$

$$L2 = \frac{\Sigma_i^B \Sigma_{x,y}^{H,W} \| (G_i[x,y] - O_i[x,y]) \|_2}{H \times W \times B} \tag{3}$$

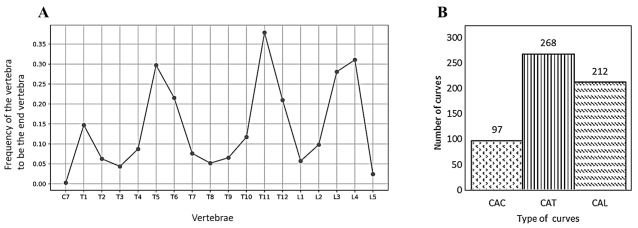


FIGURE 3. Summary of the dataset labels. Panel A indicated the frequency of each vertebra to be chosen as an end vertebra (x-axis: vertebrae; y-axis: end vertebrae frequency). B indicated the number of curves (y-axis) appeared as CAC, CAT or CAL respectively (x-axis).

III. RESULTS

In this dataset, the location of the end vertebrae (Fig. 3A) and the number of major curves (Fig. 3B) was imbalanced. End-vertebrae were more frequently identified at the 5th and the 11th thoracic vertebrae (T5 and T11), as well as the 3rd and the 4th lumbar vertebrae (L3 and L4). An increased number of curves appeared in the thoracolumbar region (CAT=268; CAL=212), compared to the cervicothoracic region (number of CAC=97). The GT CAs of the dataset ranged from 10.08° to 82.48° (average 27.55°±13.41, Table 1). Measurements of the GT CAs had an absolute inter-rater variability of 4° to 6° between two spine specialists (mean = 4.5° ± SD 0.6, ICC=0.91).

Using SpineHRNet, endplate landmarks detection was accurate in visual evaluation (Fig. 4) and the average retrieval rate (recall) in 5-fold cross-validation was 0.99 ± 0.009 (suggesting that almost all vertebral landmarks had been well retrieved). The L2 error between the predicted landmarks and the specialist-labelled ground truth landmarks was minimal, being 2.8 pixels (Fig. 4).

TABLE 1. Summary of the ground truth CAs.

Curves	Number of curves	Min	Max	Average	Standard deviation
CAC	97	12.95°	67.77°	28.98°	12.78°
CAT	268	10.73°	82.48°	29.35°	14.45°
CAL	212	10.08°	75.16°	24.7°	11.66°
CA total	577	10.08°	82.48°	27.55°	13.41°

For the predictive accuracy of the CAs generated from our newly proposed technology, the mean error of the predicted CAs was a 3.73-4.48° difference from the GT with a standard deviation of 3.11-3.64° (Table 2). The recall (0.62-0.83), precision (0.78-0.88), and F1-score (0.69 - 0.88) had the lowest predictive accuracy in the cervicothoracic curvature (Table 2: CAC) and highest accuracy in the thoracic curvature (Table 2: CAT).

The predictive reliability of the SpineHRNet based automated CAs was tested using a linear regression analysis of predicted results against the GT (Table 3, Fig. 5). The results were significantly correlated with the GT, with an overall R^2 of 0.833 and p <0.001 (Table 3). The slope of the regression line for all CAs was 42° (Table 3, Fig. 5) and close to the ideal value of 45° (indicating a perfect match between the predicted results and the GT). However, a relatively low R^2 (0.787) and regression slope (38°) was found for the CACs predictive accuracy, whereas a high R^2 (0.83) and regression slope (42°) was found during the evaluation of CATs and CALs. The overall mean difference between the GT and the predicted CAs was minimal being -0.27 (Fig. 6). Similar to the regression tests, the largest mean difference was also revealed in the reliability test of CACs (-0.62), demonstrating that agreement rate between the

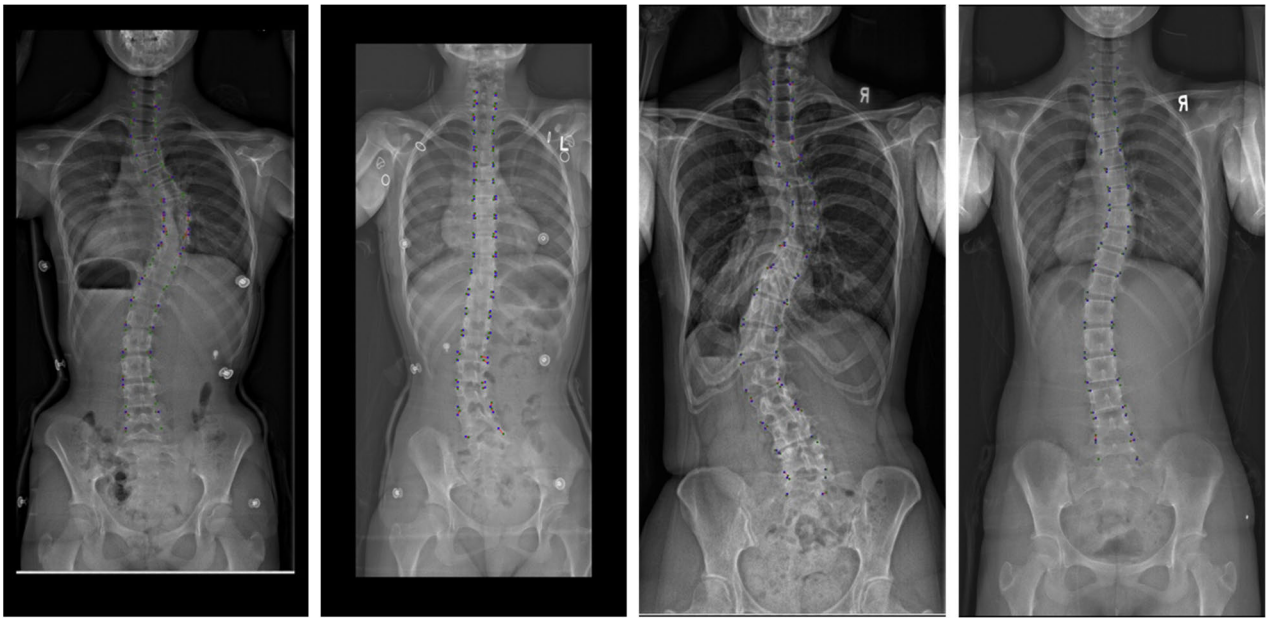


FIGURE 4. Four examples of comparison of the ground truth and the automated detections. The green points denote the ground truth (GT) landmarks and the blue points denote the predicted landmarks. Red lines connect each predicted landmark and corresponding GT landmark, demonstrating the small difference between the detection and the GT. The retrieval rate was 0.99 ± 0.009, thus the majority of the green ground truth and blue predicted points are overlapped.

TABLE 2. Evaluation metrics on CA prediction accuracy between the ground truth and the SpineHRNet predicted CAs.

Curves	Recall	Precision	F1-score	Angle error	Standard deviation
CAC	0.62	0.78	0.69	4.09°	3.35°
CAT	0.88	0.88	0.88	4.48°	3.32°
CAL	0.83	0.80	0.82	3.73°	3.64°
CA total	0.82	0.83	0.83	4.15°	3.11°

TABLE 3. Regression analysis of the correlation between ground truth CAs and those predicted by SpineHRNet.

Curves	R^2	p value	Slope of the regression line	Standard error of the measurement (S)
CAC	0.787	<0.01	38°	1.742°
CAT	0.832	<0.01	42°	0.856°
CAL	0.839	<0.01	42°	0.802°
CA total	0.833	<0.001	42°	0.558°

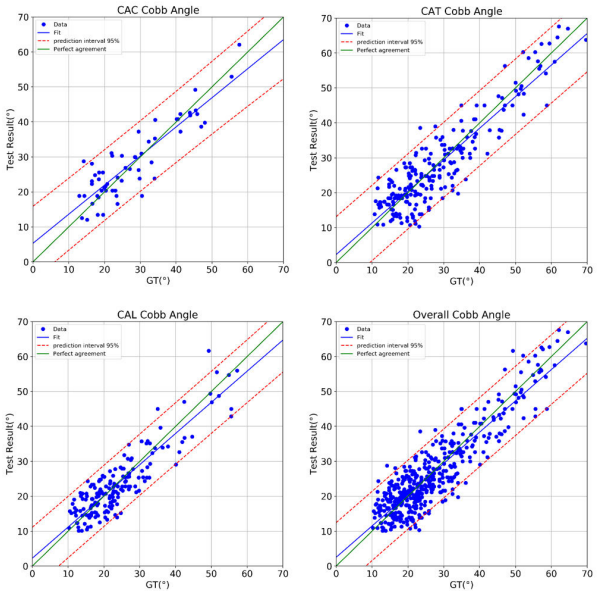


FIGURE 5. Regression analysis of the predicted alignment parameter detections (y-axis) versus the ground truth results measured manually by the spine specialists (x-axis). For CAs through the 5-folds of the predictive reliability test, good agreement between the auto-detected degrees and the ground truth was observed. All units are in degrees.

predicted results and the GT was lowest in the cervicothoracic region.

Close examination of the end vertebra predictive accuracy was also conducted (Fig. 7). From the plotted heatmap, it could be seen that despite the location of the curves (Fig. 7A&B), the end vertebrae could be accurately predicted. No false positives were presented in the test dataset (Fig. 7C). There was one interesting case of false negative (Fig. 7D) in the cervicothoracic region. However, during a close examination of the GT for this case, the CA was small at 10.73°.

IV. DISCUSSION

This is the first study to achieve accurate detection of 72 endplate landmarks and end vertebrae for C7-L5 on X-ray images despite suboptimal and variable image qualities,

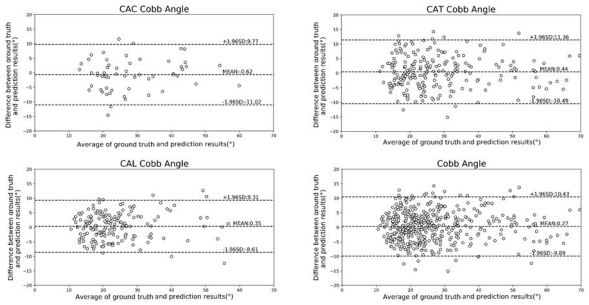


FIGURE 6. Bland-Altman plots comparing the agreement of CAs between the SpineHRNet predictions and the ground truth. The Y-axis indicates the difference between automated results and the ground truth. The X-axis represents the average of these measures ((automated results + ground truth)/2). Small mean differences from −0.62° to −0.35° with the overall mean difference of −0.27° were shown between the auto-detected CAs and the ground truth. All units are in degrees.

enabling auto-alignment for clinical analysis. While previous deep learning methods used other methods and original high-quality X-rays yielded lower accuracy with limitation of the curve patterns [24], [31]. Using our method, the CAs could be automatically determined for variable patterns of the curve. It may be due to the fact that the method we developed is suitable for the task, and the GT landmarks were labelled by spine specialists providing consistent output. To our knowledge, this is the first and largest dataset of optical images of coronal X-rays displayed on PACS for the application of HRNet to detect the key landmarks. The image size, rotation and quality variance of this dataset were large, representing real-life scenarios in telemedicine. Thus, we can foresee the application of this learning-based fully automated method in accelerating follow-up, out of hospital consultation, large scale clinical trials to avoid laborious manual assessment and inter-rater variance.

Current manual or semi-automatic alignment assessment software (including Surgimap, X-Align, Integrated Global Alignment, etc.) utilize original X-rays for spine alignment assessment. The existing software requires specialists to operate for landmark placing, whereas a system designed for fast malalignment screening without specialists' manual operations and original X-rays is not currently available. Essentially, compared with traditional manual approaches, our deep learning-based methods can be trained end-to-end in a data-driven manner and can better handle different challenges encountered to generate reproducible measurements.

Previous studies demonstrated that CAs could be computed based on the original X-Rays [10], [14], [16]–[19], [23]. One study even directly regressed CAs from input images [23]. However, learning to detect CAs as a recognition/classification task is more stable than learning to predict CAs directly as a regression task. Even trained specialists find it difficult to determine CA from an image directly without identifying the end vertebrae and measuring the slopes of their endplates using measurement tools. Thus, the application of intermediate supervision (using endplate landmarks) can enhance the reliability and the interpretability of the

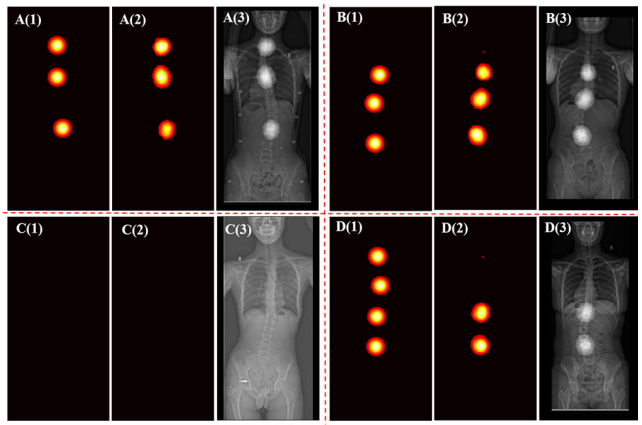


FIGURE 7. Four case examples of end vertebrae detection. A: CAC and CAT; B: CAT and CAL angle; C: normal with no curves; D: CAC and CAT and CAL angle. For each part, the ground truth heatmap (1), predicted heatmap (2) and original image merged with predicted heatmap (3) were illustrated. A false negative in the cervicothoracic region was shown in D.

predicted results. Comparably, Horng *et al.* [19] performed spine segmentation using a U-Net [32] and then computed CAs from segmentation results, which did not result in accurate detections of endplates compared to that of spine specialists. The main reason is clinically the CAs are not calculated based on the vertebral segmentation. Especially for spines with deformities, the superior or inferior borders of the segmented vertebrae are not always aligned with the endplates, while the endpoints of the endplates are essential landmarks for the CA computation. Thus, key point detections performed in our study mimicking the practice of spine surgeons, although with low-resolution and non-original X-rays, significantly improved the accuracy of the CA computation.

Other AI-integrated methods include a semi-automated algorithm for CA computation [22]. Unlike our fully automated approach, users are required to manually select several patches of end-vertebrae used to directly regress CA. Furthermore, a CNN-based network was used previously to directly regress the pixel coordinates of vertebral landmarks [21]. There are also other studies that have used multi-view X-rays as the training dataset to predict alignment parameters [24], based on original X-rays archived directly from PACS, which is difficult to obtain for telemedicine. Our approach has the advantage of being flexible and capable of handling different CA patterns while generating consistent assessment results.

The accurate detection of vertebral landmarks by our method also improved CA predictive accuracy. CA prediction (absolute mean error=3.73-4.15°, standard deviation=0.8-1.7°) was significantly more accurate than previously reported intra- and inter-rater variance (6.34-9.038°) of measurements by spine specialists using either manual or digital tools [33]. Inconsistency between specialists for CA measurements was reported with a range of 3-10° degrees resulting from different end vertebrae selection and/or manually drawing variable best-fit lines to the end vertebrae endplates [17]. This was comparable with the inter-rater variance of our specialists. By eliminating the dependency of human input, our method eliminated intra- and inter-rater variations in CA computation. Compared to a recently published conventional CNN-based deep learning method to predict CA using original X-rays [24], with a standard error of 9.9° in the predicted CA, our method generated a significantly lower error using low quality and variable aligned images, providing possibilities of applying our method to clinical practice.

The reason of this accuracy improvement is two-fold. Firstly, HRNet [25] is superior to other networks in detecting key landmarks. Secondly, unlike the majority of previous work, which formulated the vertebral landmark-detection as single-spine landmark-detection, we formulated it as multi-vertebra landmark-detection [29]. This approach was identical to the perception and workflow of spine specialists, since we captured the hierarchical relation between spine, vertebrae and endplates. Further, previous work focused on the accuracy of CA prediction, lacking examinations of intermediate output to substantiate result reliability. Endplate heatmaps serve as an intermediate supervision for interpretability, enabling specialists to evaluate directly on the reliability. It is necessary to note that consistent with the decreased number of CAC (Fig. 3B: 97) and increased number of CAT and CAL (Fig. 3B: 268&212) in our datasets, the predictive accuracy and reliability of CA in the thoracolumbar region was higher than the cervicothoracic region. Therefore, with an increased number of images collected through our future study, we expect an improved accuracy of the predicted CA. A false negative detection of the end vertebra (Fig. 7D) in the cervicothoracic region was noted, with the GT CA small at 10.73°, at the border for normal. According to the current clinical gold standard, a CA less than 10° is considered as normal, whereas between 10-20° is mild and 20-40° is moderate with a curve larger than 40° being severe.

To further justify the use of the landmark labels on re-captured radiographs as GTs, the CAs measured by specialists in the PACS (mean = 29.47°± SD 13.5) and the GT CAs calculated based on landmarks labelled on the re-captured images (Table 1) were compared and revealed no significant differences (P value = 2.0, paired t-test). Additional comparison was done between the CAs measured in the PACS and the CAs predicted through the trained SpineHRNet (mean = 27.18°± SD 13.6). No significant differences were observed between these two sets of values. Moreover, a significant linear regression association ($R^2 = 0.79$) was observed, which is similar to the R^2 between the CAs generated by GTs and SpineHRNet (Table 3).

A limitation of this study is the lack of vertebral malformation and congenital scoliosis in this dataset, as well as lack of post-operative patients with spinal instrumentation. We excluded these subjects because the number of congenital deformities was small, usually with severely deformed spine and vertebrae. To simplify the learning task, congenital and post-operative patients were not collected. A larger dataset consisting of post-operative X-rays

and congenital deformities should be established for future study.

V. CONCLUSION

Based on a collection of images of scoliosis X-rays using smartphones, a fully automatic vertebrae landmark detection and CA prediction pipeline for AIS was developed. This method can have significant clinical applications in AIS screening, follow-up, as well as facilitating deformity research by providing accurate and mobile CA detections for telemedicine, reducing specialists' burden for radiographic measurements with increased assessment consistency.

ACKNOWLEDGMENT

The authors would like to thank their LabTechnician Mr. Huiqian Zhou and their Student Intern Ms. Trixie Mak for organizing the captured images. They sincerely appreciate Prof. Ashish Diwan on advising this study. *(Teng Zhang and Yifei Li contributed equally to this work.)*

REFERENCES

[1] D. Y. Fong, K. M. Cheung, Y. W. Wong, Y. Y. Wan, C. F. Lee, T. P. Lam, J. C. Cheng, B. K. Ng, and K. D. Luk, "A population-based cohort study of 394,401 children followed for 10 years exhibits sustained effectiveness of scoliosis screening," *Spine J.*, vol. 15, no. 5, pp. 825–833, May 2015, doi: 10.1016/j.spinee.2015.01.019.

[2] M. de Seze and E. Cugy, "Pathogenesis of idiopathic scoliosis: A review," *Ann. Phys. Rehabil. Med.*, vol. 55, no. 2, pp. 38–128, Mar. 2012, doi: 10.1016/j.rehab.2012.01.003.

[3] N. Chung, Y.-H. Cheng, H.-L. Po, W.-K. Ng, K.-C. Cheung, H.-Y. Yung, and Y.-M. Lai, "Spinal phantom comparability study of cobb angle measurement of scoliosis using digital radiographic imaging," *J. Orthopaedic Transl.*, vol. 15, pp. 81–90, Oct. 2018, doi: 10.1016/j.jot.2018.09.005.

[4] A. Y. L. Wong, D. Samartzis, P. W. H. Cheung, and J. P. Y. Cheung, "How common is back pain and what biopsychosocial factors are associated with back pain in patients with adolescent idiopathic scoliosis?" *Clin. Orthopaedics Rel. Res.*, vol. 477, no. 4, pp. 676–686, Apr. 2019, doi: 10.1097/CORR.0000000000000569.

[5] L. E. Peterson and A. L. Nachemson, "Prediction of progression of the curve in girls who have adolescent idiopathic scoliosis of moderate severity. Logistic regression analysis based on data from the brace study of the scoliosis research Society.," *J. Bone Joint Surg.*, vol. 77, no. 6, pp. 823–827, Jun. 1995, doi: 10.2106/00004623-199506000-00002.

[6] J. E. Lonstein and J. M. Carlson, "The prediction of curve progression in untreated idiopathic scoliosis during growth.," *J. Bone Joint Surg.*, vol. 66, no. 7, pp. 1061–1071, Sep. 1984, doi: 10.2106/00004623-198466070-00013.

[7] S. L. Weinstein, L. A. Dolan, J. G. Wright, and M. B. Dobbs, "Effects of bracing in adolescents with idiopathic scoliosis," *New England J. Med.*, vol. 369, no. 16, pp. 1512–1521, Oct. 2013, doi: 10.1056/NEJMoa1307337.

[8] X. Sun, Q. Ding, S. Sha, S. Mao, F. Zhu, Z. Zhu, B. Qian, B. Wang, J. C. Y. Cheng, and Y. Qiu, "Rib-vertebral angle measurements predict brace treatment outcome in risser grade 0 and premenarchal girls with adolescent idiopathic scoliosis," *Eur. Spine J.*, vol. 25, no. 10, pp. 3088–3094, Oct. 2016, doi: 10.1007/s00586-015-4372-5.

[9] A. L. Nachemson and L. E. Peterson, "Effectiveness of treatment with a brace in girls who have adolescent idiopathic scoliosis. A prospective, controlled study based on data from the brace study of the scoliosis research Society.," *J. Bone Joint Surg.*, vol. 77, no. 6, pp. 815–822, Jun. 1995, doi: 10.2106/00004623-199506000-00001.

[10] A. Safari, H. Parsaei, A. Zamani, and B. Pourabbas, "A semi-automatic algorithm for estimating cobb angle," *J. Biomed. Phys. Eng.*, vol. 9, no. 3Jun, pp. 317–326, Jun. 2019, doi: 10.31661/jbpe.v9i3Jun.730.

[11] O. A. Okashi, H. Du, and H. Al-Assam, "Automatic spine curvature estimation from X-ray images of a mouse model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 175–184, Mar. 2017, doi: 10.1016/j.cmpb.2016.12.010.

[12] J. Mukherjee, R. Kundu, and A. Chakrabarti, "Variability of Cobb angle measurement from digital X-ray image based on different de-noising techniques," *Int. J. Biomed. Eng. Technol.*, vol. 16, no. 2, pp. 113–134, 2014, doi: 10.1504/Ijbet.2014.065656.

[13] H. Anitha, A. K. Karunakar, and K. V. N. Dinesh, "Automatic extraction of vertebral endplates from scoliotic radiographs using customized filter," *Biomed. Eng. Lett.*, vol. 4, no. 2, pp. 158–165, Jun. 2014, doi: 10.1007/s13534-014-0129-z.

[14] T. A. Sardjono, M. H. F. Wilkinson, A. G. Veldhuizen, P. M. A. van Ooijen, K. E. Purnama, and G. J. Verkerke, "Automatic cobb angle determination from radiographic images," *Spine*, vol. 38, no. 20, pp. E1256–E1262, Sep. 2013, doi: 10.1097/BRS.0b013e3182a0c7c3.

[15] A. H and G. K. Prabhu, "Automatic quantification of spinal curvature in scoliotic radiograph using image processing," *J. Med. Syst.*, vol. 36, no. 3, pp. 1943–1951, Jun. 2012, doi: 10.1007/s10916-011-9654-9.

[16] R. Kundu, A. Chakrabarti, and P. K. Lenka, "Cobb angle measurement of scoliosis with reduced variability," 2012, *arXiv:1211.5355*. [Online]. Available: http://arxiv.org/abs/1211.5355

[17] J. Zhang, E. Lou, X. Shi, Y. Wang, D. L. Hill, J. V. Raso, L. H. Le, and L. Lv, "A computer-aided Cobb angle measurement method and its reliability," *J. Spinal Disorders Techn.*, vol. 23, no. 6, pp. 383–387, Aug. 2010, doi: 10.1097/BSD.0b013e3181bb9a3c.

[18] J. Zhang, E. Lou, L. H. Le, D. L. Hill, J. V. Raso, and Y. Wang, "Automatic Cobb measurement of scoliosis based on fuzzy Hough Transform with vertebral shape prior," *J. Digit. Imag.*, vol. 22, no. 5, pp. 463–472, Oct. 2009, doi: 10.1007/s10278-008-9127-y.

[19] M.-H. Horng, C.-P. Kuok, M.-J. Fu, C.-J. Lin, and Y.-N. Sun, "Cobb angle measurement of spine from X-ray images using convolutional neural network," *Comput. Math. Methods Med.*, vol. 2019, Feb. 2019, Art. no. 6357171.

[20] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, "Automated comprehensive adolescent idiopathic scoliosis assessment using MVC-net," *Med. Image Anal.*, vol. 48, pp. 1–11, Aug. 2018, doi: 10.1016/j.media.2018.05.005.

[21] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, "Automatic landmark estimation for adolescent idiopathic scoliosis assessment using boostnet," presented at the MICCAI, Sep. 2017.

[22] J. Zhang, H. Li, L. Lv, and Y. Zhang, "Computer-aided cobb measurement based on automatic detection of vertebral slopes using deep neural network," *Int. J. Biomed. Imag.*, vol. 2017, pp. 1–6, Oct. 2017, doi: 10.1155/2017/9083916.

[23] H. Sun, X. Zhen, C. Bailey, P. Rasoulinejad, Y. Yin, and S. Li, "Direct estimation of spinal Cobb angles by structured multi-output regression," in *Proc. Int. Conf Inf. Process. Med. Imag.*, in (Lecture Notes in Computer Science), vol. 10265. Boone, NC, USA, Springer, 2017, pp. 529–540.

[24] F. Galbusera, F. Niemeyer, H.-J. Wilke, T. Bassani, G. Casaroli, C. Anania, F. Costa, M. Brayda-Bruno, and L. M. Sconfienza, "Fully automated radiological analysis of spinal disorders and deformities: A deep learning approach," *Eur. Spine J.*, vol. 28, no. 5, pp. 951–960, May 2019, doi: 10.1007/s00586-019-05944-z.

[25] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.

[26] R. Swinfen and P. Swinfen, "Low-cost telemedicine in the developing world," *J. Telemed. Telecare*, vol. 8, no. 3, pp. 63–65, Dec. 2002.

[27] E. Ozdalga, A. Ozdalga, and N. Ahuja, "The smartphone in medicine: A review of current and potential use among physicians and students," *J. Med. Internet Res.*, vol. 14, no. 5, p. e128, Sep. 2012.

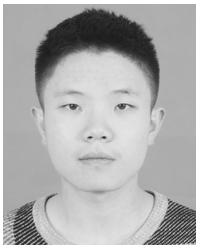
[28] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2014, pp. 1799–1807. [Online]. Available: Go to ISI>://WOS:000452647103024

[29] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. 31th Int. Conf. Neural Inf. Process. Syst.*, vol. 30, Jan. 2017, pp. 2274–2284. [Online]. Available: Go to ISI>://WOS:000452649402032

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] L. Wang, Q. Xu, S. Leung, J. Chung, B. Chen, and S. Li, "Accurate automated cobb angles estimation using multi-view extrapolation net," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101542, doi: 10.1016/j.media.2019.101542.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," presented at the MICCAI, Oct. 2015.
[33] M. Gstoettner, K. Sekyra, N. Walochnik, P. Winter, R. Wachter, and C. M. Bach, "Inter- and intraobserver reliability assessment of the cobb angle: Manual versus digital measurement tools," *Eur. Spine J.*, vol. 16, no. 10, pp. 1587–1592, Oct. 2007, doi: 10.1007/s00586-007-0401-3.

TENG ZHANG (Member, IEEE) received the B.Med.Sc. and M.B.M.E. degrees from the University of New South Wales, Sydney, Australia, in 2010, and the Ph.D. degree in medicine from the University of New South Wales, in 2016. From 2011 to 2017, she worked with the St George Hospital, Sydney, as a Scientific Officer, where she has collaborated with several medical device companies in system optimization and clinical validations. Since 2018, she has been with the Department of Orthopaedics and Traumatology, The University of Hong Kong, where she currently serves as a Research Officer. She has authored over thirty peer-reviewed articles with over three hundred citations. Her research interest includes modeling of biological systems with direct clinical applications use both conventional and learning based methods.



YIFEI LI received the B.E. degree in computer science and technology from Zhejiang University, Zhejiang, China, in 2020. He was a Research Assistant with the Laboratory of Computer Vision, The University of Hong Kong, China. His research interests include the system and machine learning, which applies machine learning algorithms to tune the system performance and designs system to support large-scale machine learning.

JASON PUI YIN CHEUNG received the M.B.B.S. and Master of Medical Sciences degrees from The University of Hong Kong, in 2007 and 2012, respectively, the Master of Surgery degree, in 2017, the Postgraduate Diploma degree in molecular and diagnostic pathology, in 2018, and the Doctor of Medicine degree, in 2019. In November 2012, he joined the Department of Orthopaedics and Traumatology, as a Clinical Assistant Professor, and promoted to a Clinical Associate Professor with early tenure, in 2018. He completed training in orthopaedics at Queen Mary Hospital and specialist training, in 2014. He has authored over 180 peer-reviewed articles with over 1600 citations. His main research interests include paediatric growth and spinal deformity, and developmental lumbar spinal stenosis.

SOCRATES DOKOS (Member, IEEE) is currently an Associate Professor with the Graduate School of Biomedical Engineering, University of New South Wales, Sydney, Australia. He has authored over one hundred and sixty peer reviewed journal articles, book chapters and conference proceedings, and has collaborated with several medical device companies using computational modeling to better understand and improve their device performance. He has also authored a COMSOL-based book *Modelling Organs, Tissues, Cells and Devices: Using Matlab and COMSOL Multiphysics* (Springer, 2017), which has achieved in excess of 31K individual chapter downloads worldwide. His research interests include the development of computational models for various biomedical engineering applications, including neurostimulators, visual prostheses, transcranial electric stimulators, cardiac defibrillators, left ventricular assist devices, artificial mitral valves, and other biomechanics applications.

KWAN-YEE K. WONG (Senior Member, IEEE) received the B.Eng. degree (Hons.) in computer engineering from The Chinese University of Hong Kong, in 1998, and the M.Phil. and Ph.D. degrees from the University of Cambridge, in 2000 and 2001, respectively, both in computer vision (information engineering). Since 2001, he has been with the Department of Computer Science, The University of Hong Kong, where he is currently an Associate Professor. His research interests include computer vision and machine intelligence. He is also an Editorial Board Member of *International Journal of Computer Vision (IJCV)*.