Radiograph-comparable image synthesis for spine alignment analysis using deep learning with prospective clinical validation

Nan Meng,[a,c] Kwan-Yee K. Wong,[b] Moxin Zhao,[a] Jason P. Y. Cheung,[a,*] and Teng Zhang[a,c,**]

[a]Department of Orthopaedics and Traumatology, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
[b]Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
[c]CoNova Medical Technology Limited, Hong Kong SAR, China

eClinicalMedicine
2023;61: 102050

Published Online xxx
https://doi.org/10.1016/j.eclinm.2023.102050

Summary

Background Adolescent idiopathic scoliosis (AIS) is the most common type of spinal disorder affecting children. Clinical screening and diagnosis require physical and radiographic examinations, which are either subjective or increase radiation exposure. We therefore developed and validated a radiation-free portable system and device utilising light-based depth sensing and deep learning technologies to analyse AIS by landmark detection and image synthesis.

Methods Consecutive patients with AIS attending two local scoliosis clinics in Hong Kong between October 9, 2019, and May 21, 2022, were recruited. Patients were excluded if they had psychological and/or systematic neural disorders that could influence the compliance of the study and/or the mobility of the patients. For each participant, a Red Green Blue-Depth (RGBD) image of the nude back was collected using our in-house radiation-free device. Manually labelled landmarks and alignment parameters by our spine surgeons were considered as the ground truth (GT). Images from training and internal validation cohorts (n = 1936) were used to develop the deep learning models. The model was then prospectively validated on another cohort (n = 302) which was collected in Hong Kong and had the same demographic properties as the training cohort. We evaluated the prediction accuracy of the model on nude back landmark detection as well as the performance on radiograph-comparable image (RCI) synthesis. The obtained RCIs contain sufficient anatomical information that can quantify disease severities and curve types.

Findings Our model had a consistently high accuracy in predicting the nude back anatomical landmarks with a less than 4-pixel error regarding the mean Euclidian and Manhattan distance. The synthesized RCI for AIS severity classification achieved a sensitivity and negative predictive value of over 0.909 and 0.933, and the performance for curve type classification was 0.974 and 0.908, with spine specialists' manual assessment results on real radiographs as GT. The estimated Cobb angle from synthesized RCIs had a strong correlation with the GT angles ($R^2 = 0.984$, $p < 0.001$).

Interpretation The radiation-free medical device powered by depth sensing and deep learning techniques can provide instantaneous and harmless spine alignment analysis which has the potential for integration into routine screening for adolescents.

Funding Innovation and Technology Fund (MRP/038/20X), Health Services Research Fund (HMRF) 08192266.

Keywords: Adolescent idiopathic scoliosis; Radiation-free; Light-based; Radiograph-comparable image synthesis; Deep learning

Introduction

Scoliosis is a three-dimensional (3D) deformity of the spine defined as a Cobb angle[1] (CA: measured by an angle formed by the upper endplate of the uppermost tilted vertebra and the lower endplate of the lowermost tilted vertebra of the structural curve) greater than 10°

Abbreviations: AIS, adolescent idiopathic scoliosis; GT, ground truth; RCI, radiograph-comparable image; CA, Cobb angle; MIS, medical image synthesis

*Corresponding author. 5/F, Professorial Block, Queen Mary Hospital, Pokfulam, Hong Kong SAR, China.
**Corresponding author. Department of Orthopaedics and Traumatology, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.
E-mail addresses: cheungjp@hku.hk (J.P.Y. Cheung), tgzhang@hku.hk (T. Zhang).

Research in context

Evidence before this study

PubMed was searched on August 29, 2022 for all research articles containing the terms "artificial intelligence" OR "deep learning" AND "radiation-free" OR "no-radiation" AND "scoliosis", without any date or language restrictions. Cited references in the retrieved articles were further searched. Only six relevant studies were identified that explored radiation-free techniques based on ultrasound images, back surface point clouds and RGB images as medical data for spine alignment analysis. Most of these studies suffered from data scarcity and were validated in small cohorts without prospective validation. To date, no study has explored the potential of deep learning for medical image synthesis (MIS) using non-medical imaging modality data.

Added value of this study

To our knowledge, this study is the first to develop a non-radiation medical device powered by deep learning models for accurate anatomical landmark detection and realistic radiographic-comparable images (RCIs) synthesis using optical and depth images. Furthermore, the synthesised RCIs can be used to estimate the Cobb angle and the estimated results have excellent agreement and correlation with the ones calculated from real radiographs. Our work contributes to the potential usage of fast and harmless assessments of scoliosis in clinics, underscores the value of light-based depth sensing techniques in scoliosis diagnosis and treatment, and provides insight into a new direction of research for non-radiation bone alignment assessment. Our model achieved satisfactory clinical performance with prospective evaluation.

Implications of all the available evidence

A non-radiation device with the deep learning powered analysis platform can provide rapid and harmless examination of spine and has the potential for integration into routine screening of adolescents. Our study demonstrates a promising performance of using light-based depth sensing techniques in conjunction with deep learning for scoliosis analysis. This screening tool may be of use in resource-constrained or remote regions, such as those with a shortage of radiographic medical imaging devices or specialists. An additional international multi-centre trial is needed to assess the impacts of demographic variables, such as body-mass index and skin colour, and improve the reliability of our system and device before clinical use.

on standing plain radiographs. Among all types of scoliosis (i.e., idiopathic, congenital, neuromuscular, and syndromic), adolescent idiopathic scoliosis (AIS) is most common in the paediatric population, with a prevalence up to 2.2% of boys and 4.8% of girls[2,3] in Hong Kong. Untreated cases may progress rapidly during the pubertal growth spurt, causing body disfigurement, cardiopulmonary compromise, and back pain.[4,5] In addition, the degeneration of the spine may damage the surrounding muscles, ligaments and joint structures which can worsen pain and cause additional physical limitations. Early detection and interventions to prevent curve progression is therefore critical.[6]

Screening using forward bending assessment and scoliometer measurement are the main options to identify individuals with AIS.[7,8] Radiographic examination is the reference standard to quantify AIS severity and curve types.[9] AIS follow-up and progression monitoring require repetitive radiographic examinations.[10] Children are especially sensitive to radiation due to higher metabolic activity of their cells.[11] Thus, accurate and radiation-free approaches[12] for spine alignment analysis is desirable.

Non-radiation techniques for scoliosis assessment have been studied for years. Previous methods include 3-dimensional ultrasound,[13] digital inclinometer,[14] rasterstereography[15] and electrogoniometer.[16] Deep learning has made considerable progress in image generation and transformation.[17,18] Clinically, such techniques have been used for medical image synthesis (MIS) to facilitate the clinical workflow,[19,20] for example, treatment planning[21] and PET attenuation correction.[22] Despite the diverse applications, most of the studies focused on image synthesis between two medical modalities,[23,24] and seldom explored the feasibility of image synthesis between medical and optical imaging systems.

We aim to accurately quantify AIS spine malalignment with no radiation. Thus, Light-Based *Radiograph-Comparable Image (RCI) Synthesis* was explored, and a deep learning approach to synthesize RCI from nude back images containing RGB and depth information (RGBD) was developed (Fig. 1a). Different from the previous techniques, our models generate synthesized RCI containing anatomical morphology information of the spine to accurately quantify spinal alignment. We prospectively validated the reliability of our models with multiple tasks in two clinics, including back landmark auto-detection, RCI synthesis, and scoliosis severity and curve type classification. Our technology has the potential to facilitate radiation-free, fast, and accurate AIS analysis.

Methods

Data collection and preparation

From October 9, 2019, to January 15, 2022, all consecutive patients with AIS aged between 10 and 18 years old were recruited to form the training and internal validation dataset for the technology development. From January 16, 2022, to May 21, 2022, consecutive patients with AIS who attended two local clinics in Hong Kong were recruited in our prospective testing dataset. Patients were excluded if they had psychological and/or systematic

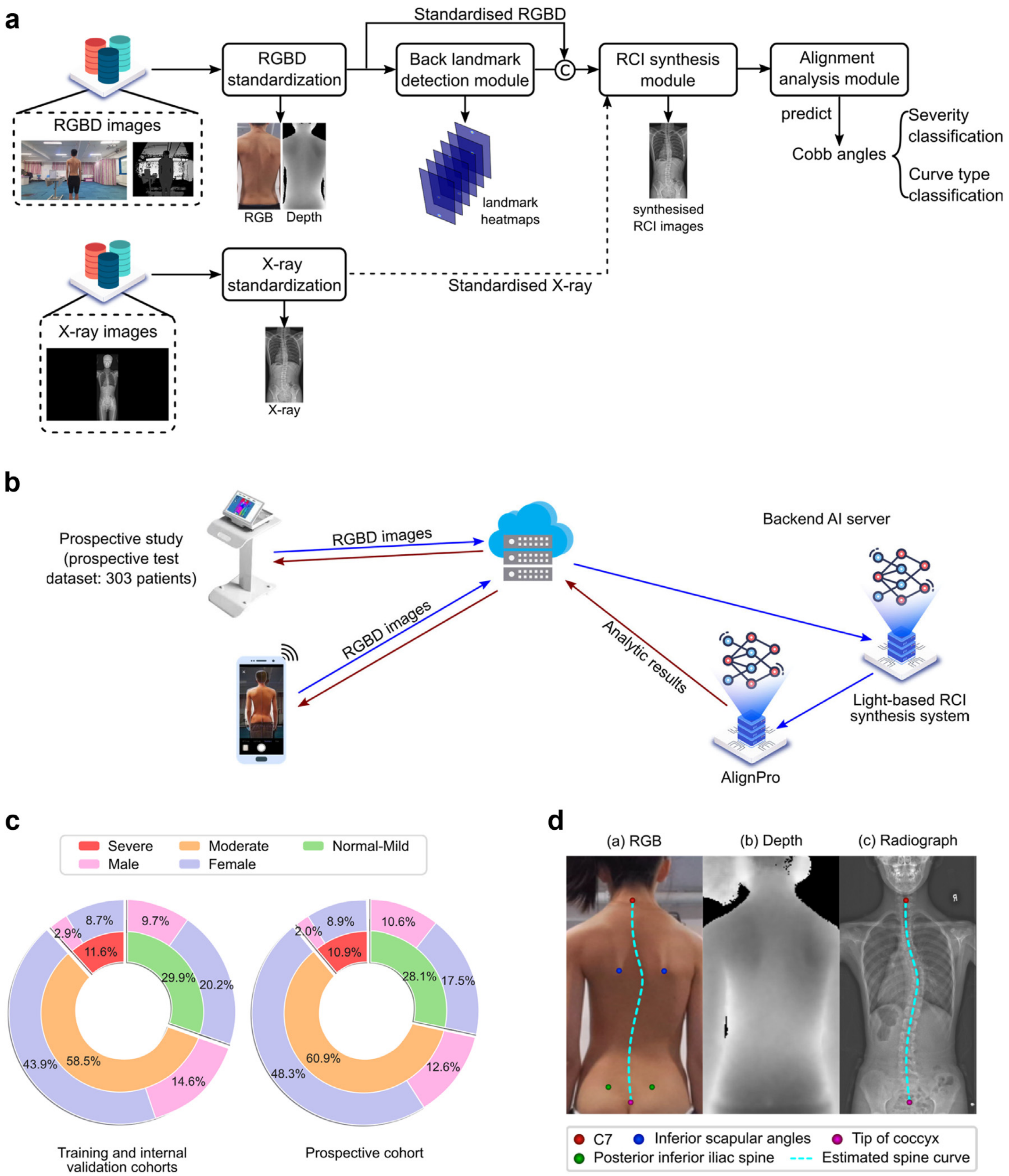

Fig. 1: **The pipeline of light-based RCI synthesis system and the demographics of data**. **a**, Illustration of the pipeline of the light-based RCI synthesis system. The system includes a pipeline consisting of 1) a RGBD and radiograph standardization module, 2) a back landmark detection module, 3) a landmark guided RCI synthesis module and 4) a quantitative alignment analysis module. The first is implemented with rule-based and adaptive algorithms to standardize the images, while the last three modules adopt deep learning techniques. **b**, Application and evaluation of the AI system. RGBD images captured with the smartphone and our equipment are transmitted to the cloud data center and backend AI server that hosting the light-based RCI synthesis and AlignPro system for analysis. Then, the results can be instantly transmitted and displayed back to the smartphone and equipment. **c**, Proportion of different severity levels and genders in all participants. **d**, An example of aligned RGB, depth and corresponding radiograph. Abbreviation definition: RCI: Radiograph-comparable image; RGB: Red green blue; RGBD: Red green blue-depth; AI: Artificial intelligence.

neural disorders that could influence the compliance of the study and/or the mobility of the patients. Other exclusion criteria were: 1) any trauma that might impair posture and mobility, 2) severe dermatological conditions that might impair the optical imaging results, and 3) any known oncological disease. The dataset included patients from two scoliosis clinics: 1) The Duchess of Kent Children's Hospital and 2) Queen Mary Hospital in Hong Kong. The sex of each participant was determined in terms of physiological characteristics (according to the information on his/her ID card). For each patient, a bareback RGBD image and a whole-spine standing posteroanterior radiograph were acquired. The RGBD imaging system consists of an RGBD camera,[25] a portable computer and a self-designed mobile stand (Supplementary Fig. S1 and Supplementary Fig. S2). The radiograph was acquired using the EOS™ (EOS® Imaging, Paris, France) biplanar stereoradiography machine. The differences between the two clinics are minimal and a detailed study protocol was prepared to guide the experimental settings and patient behaviours when collecting data. To avoid any misoperation, the involved technicians and clinical assistant were well-trained before joining the study. The proposed system consists of four modules (Fig. 1a): 1) a RGBD and radiograph standardization module, 2) a back landmark detection module, 3) a landmark guided RCI synthesis module, and 4) a quantitative alignment analysis module.

Ethics

Institutional authority review board (UW15-596) approval was obtained in the two local clinics in Hong Kong and all participants were required to provide written informed consent before joining this study.

Image pre-processing

Image pre-processing was conducted to standardize the images and resolve the misalignment of RGBD images and radiographs obtained from different imaging systems. A novel image registration algorithm was developed for this purpose to align the RGBD and radiographs according to the landmarks of the 7th cervical vertebra (C7) and the tip of coccyx (TOC) in radiographs and RGBD images. To eliminate the capture angle impacts when collecting RGBD images, random sample consensus (RANSAC) was used to standardize the image capture angle. Each captured image was standardized according to the checkboard plane which was placed perpendicular to the ground and beside the patient (as shown in Supplementary Fig. S3a). Finally, both aligned radiograph and RGBD image were cropped to a 512 × 256 patch, only containing the back region (as shown in Supplementary Fig. S3b).

Data labelling

For RGBD images, 6 anatomical landmarks on the nude back were annotated according to experienced surgeons' recommendations, including 1) C7, 2) left inferior scapular angle, 3) right inferior scapular angles, 4) left posterior inferior iliac spine (PIIS), 5) right PIIS, and 6) TOC (Fig. 1d). For radiographs, two landmarks (i.e., C7 and TOC) were annotated as the registration reference (Fig. 1d). All landmarks were manually annotated by our spine surgeons with more than 20 years of clinical experience. The CA of each radiograph was manually labelled by two spine specialists. Measurements of the CAs had an absolute inter-rater variability from 4 to 6 (mean = 4.5 ± SD 0.6) between the two spine specialists.[20]

Definition of deformity severity, curve types

The severity of spine deformity was classified into 3 categories according to our clinical standard.[26,27] Curves with a CA larger than 40° were classified as severe, between 20° and 40° were considered as moderate, between 0° and 20° were considered as normal-mild. The deformity severity assessment standard and clinical management are presented in Supplementary Table S1. The curve type was regarded as thoracic (T) if the curve apex was between the 1st and the 11th thoracic vertebrae and considered as thoracolumbar or lumbar (TL/L) if the curve apex was between the 12th thoracic vertebra and the 5th lumbar vertebrae.

Light-based RCI synthesis system

The light-based RCI synthesis system consisted of a back landmark detection module and an RCI synthesis module (Supplementary Fig. S4). The back landmark detection module adopted the modified HRNet backbone (Supplementary Fig. S5) to predict the 6 anatomical landmarks. It utilised different branches to extract multi-scale features from the images and then integrated these features to achieve better model outputs. The number of channels output by the module was 6 and each channel was a heatmap, providing the probability of the position of the landmark. The RCI synthesis module adopted the CycleGAN framework but used ResNet as the generator. It took the concatenation of the RGBD images (4-channels) and the 6 detected landmark heatmaps (6-channels), in total 10-channel data, as input and outputs RCI image. The details of the ResNet model and PatchGAN model can be found in Supplementary Fig. S6.

The quantitative alignment analysis module utilised the online platform (AlignPro[27]) for automatic CA prediction. Both original radiographs and generated RCIs were sent to the server of AlignPro and then we could obtain the endplate landmarks of the end vertebrae located in different spinal curves. The obtained landmarks were further analysed and adjusted by our senior clinicians to eliminate noise in the automatic predictions and fill the missing landmarks without reference to the original radiographs. The final output CAs from this module were calculated according to the landmarks reviewed by the clinicians using the AlignPro platform.

RGBD imaging device

The RGBD imaging device mainly consists of an RGBD camera, a portable computer, and a self-designed mobile stand. The RGBD camera is an Azure Kinect DK camera which is used to record both appearance and depth information of the nude back. A portable computer (with Windows OS, an Intel Core i5 CPU and 16 GB memory) is connected to the camera and users can operate the camera for RGBD image capturing and archiving through our self-developed software. Both camera and computer are installed on a portable stand for the convenience of users. The appearance of our device for data collection is shown in Supplementary Fig. S1.

Performance metrics for landmark detection

Mean Euclidean distance (MED) and mean Manhattan distance (MMD) were adopted as quantitative measurements to evaluate the performance of the landmark detection. MED measures the average Euclidean distance between the predicted landmark and GT landmark, while MMD also measures the average distance along the axes. The definition of these two measurements is:

$$\mathrm{MED} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}\right), \tag{1}$$

$$\mathrm{MMD} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{i=1}^{M}\left(|x_i - \hat{x}_i| + |y_i - \hat{y}_i|\right)\right), \tag{2}$$

where denotes the coordinates of the GT landmark, denotes the coordinates of the predicted landmark, M is the number of landmarks and N is the patient number.

Quality metrics for the RCI synthesis

Five image quality metrics, namely Fréchet inception distance (FID),[28] Visual information fidelity (VIF),[29] Learned perceptual image patch similarity (LPIPS),[30] Natural image quality evaluator (NIQE),[31] and Blind/referenceless image spatial quality evaluator (BRISQUE),[32] were selected to evaluate the RCI synthesis performance (details refer to Supplementary Section 2.2).

Statistical analysis

To evaluate the performance of our model on back landmark detection, we analysed the statistics and error distribution for each landmark using violin plot, box plot, and scatter plot in Fig. 2. The error was measured using MED and MMD. The box plot indicates the 1st, 2nd, and 3rd quartile (Q1, Q2 and Q3), interquartile range (IQR), minimum and maximum value (exclude the outliers) of the data points. The violin plot provides an intuitive view of the error distribution. The region smaller than zero was cropped due to the nonnegativity of MED and MMD. Meanwhile, the region larger than the 95th percentile was also cropped to ignore the outliers. The scatter plot shows how the error data points distributed along the y-axis.

To quantitatively assess the validity of the CA parameters estimated from the synthesized RCI, linear regression and Bland–Altman analysis were conducted. In the regression plot (Fig. 4a), the regression line (blue line), 95% confidence interval line (green dashed line) of the predictions, and the ideal correspondence (red line) between the predicted CAs on synthesized RCIs and GT radiographs are presented. The Bland–Altman analysis was performed between the mean of the predicted CAs on RCIs and GT radiographs and their residual to assess the agreement between them.

Three classification tasks, namely severity grading (3 types: normal-moderate, moderate, and severe), curve detection, and the curve type prediction (2 types: T and TL/L), were assessed using confusion matrix analysis (Fig. 4c–f). For curve detection, we distinguished the patients from healthy controls according to the normal range of CA (CA <10°). To evaluate the severity grading and curve type classification performance of our deep learning model, five statistical measurements were calculated, including sensitivity (Sn), specificity (Sp), precision (Pr), negative predictive value (NPV), and accuracy (Acc) according to the following equations:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{3}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}, \tag{4}$$

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{5}$$

$$\mathrm{NPV} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{6}$$

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}, \tag{7}$$

where TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative predictions, respectively. All statistical analysis were done using Python (v3.8) and several python packages, including

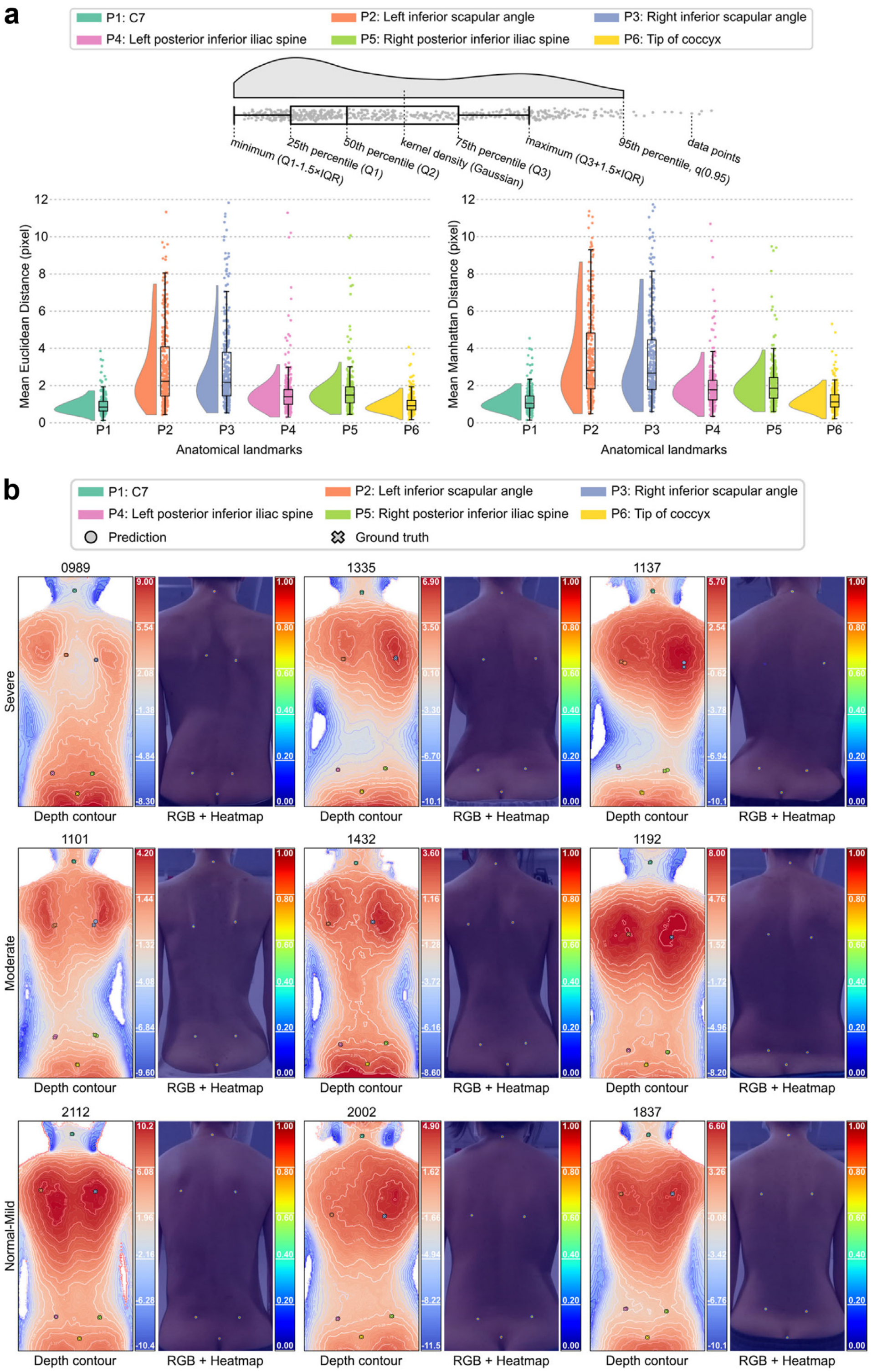

Fig. 2: **Statistical analysis and visual results of the landmarks detection on the bareback**. **a**, A combination of violin plot, box plot and scatter plot of the MED (left figure) and MMD (right figure) values for the 6 landmarks. **b**, Depth and combined heatmaps of the 6 bareback anatomical landmarks for landmark detection. Here, we present 3 samples for each disease severity (normal-mild, moderate and severe). Left side presents the contour plot of the depth image of each patient with predicted and GT landmarks. The colourbar demonstrates the height of the surface measured in millimeter in terms of the height of landmark C7. The higher the region is, the closer the region to the camera (e.g., red

Severity	Male	Female	Mean max CA, degree	Mean age (SD), years	Mean BMI (SD), kg/m^2	Curve type		
						T	TL/L	Mixed
Training and internal validation cohorts								
Normal-Mild	187	392	15.1	13.9 (2.2)	18.4 (2.7)	81	151	290
Moderate	282	850	27.5	14.5 (2.5)	18.7 (2.9)	100	222	810
Severe	57	168	49.6	15.0 (2.1)	18.8 (3.0)	24	23	178
Prospective cohort								
Normal-Mild	32	53	14.9	13.8 (2.1)	17.2 (2.2)	19	24	33
Moderate	38	146	26.5	14.5 (2.3)	17.9 (2.4)	17	44	123
Severe	6	27	46.2	16.1 (2.3)	18.4 (2.4)	4	3	26

Abbreviation definition: CA: Cobb angle; SD: Standard deviation; BMI: Body mass index; T: Thoracic; TL/L: Thoracolumbar/Lumbar.

Table 1: **Demographic information of the study population.**

Numpy (v1.18.5), SciPy (v1.5.2), Ptitprince (v0.2.6), pandas (v1.1.3), seaborn (v0.11.0), and Matplotlib (v3.3.2).

The sub-sampling sample size determination method (SSDM) was used to examine how sample size affects our two models. Unfortunately, SSDM's application to deep learning in medical imaging is still in its early stages, and there is no suitable SSDM to the problem studied in this research. As a result, we chose a practical SSDM, i.e., a curve-fitting method,[33] in this paper to empirically assess the model's effectiveness at various sample size proportions. The training data was randomly sub-sampled by a proportion factor of 4%, 8%, 16%, 32%, 64%, 100%, and for each factor, the models were trained 10 times (Supplementary Fig. S8 and Supplementary Fig. S9). For each training run, the weights from the training history with the minimal validation loss were stored and then assessed on the prospective testing dataset, yielding 10 test loss estimates for each proportion factor (Supplementary Fig. S10). For both deep learning models, with the increasing training samples, the model performance improved. For the landmark detection model, sampling proportion should be larger than 32% to ensure the model performance and stability while for the RCI synthesis model, all training samples should be used to achieve the best model performance.

Role of the funding source

Our study was funded by the Innovation and Technology Fund (MRP/038/20X) and Health Services Research Fund 08192266. The funders of the study had no role in study design, model development, data collection and analysis, interpretation, or writing of the article. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Datasets

Between October 9, 2019, and May 21, 2022, 2238 participants were enrolled. The demographic information of the technology development and prospective testing cohorts is presented in Table 1. The study population included 1936 patients (1410 female, 72.8%) in the training and internal validation cohorts for the model development. An additional 302 patients (226 female, 74.8%) were recruited prospectively from two local spine clinics for performance testing. There was no patient overlap between the model development and prospective testing cohorts. In the training and validation cohort, there were 579 patients (29.9%) classified as normal-mild, 1132 patients (58.5%) classified as moderate, and 228 patients (11.6%) classified as severe. In the prospective cohort, there were 85 patients (28.1%) classified as normal-mild, 184 patients (60.9%) classified as moderate, and 33 patients (10.9%) classified as severe. The characteristics of different cohorts are summarized in Table 1. The percentage of patients with different severities, and for each severity, the proportion for each sex is also presented (Fig. 1c).

Performance of back landmark detection on the back images

Table 2 evaluates landmark detection on the prospective test dataset. For each anatomical landmark, we reported both mean and standard deviation (SD). The detection

means closer to the camera while blue means away from the camera). Right side displays the RGB image of the patient as background, and the landmark heatmap as foreground. The colourbar denotes the probability of the appearance of the landmarks (e.g., red means higher probability and blue means lower probability). Abbreviation definition: MED: Mean euclidean distance; MMD: Mean manhattan distance; GT: Ground truth; C7: the 7th cervical vertebra. The violin plot shows the distribution of the error values. The box plot is defined as a graphical representation of the distribution of the data, with the box representing the interquartile range, the line inside the box denoting the median, and the whiskers representing the range of data.

Metrics	C7 (P1)		Left inferior scapular angle (P2)		Right inferior scapular angle (P3)		Left PIIS (P4)		Right PIIS (P5)		TOC (P6)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MED (pixels)	1.0	0.5	3.0	2.1	2.9	2.3	1.6	1.2	1.7	1.2	1.0	0.5
MMD (pixels)	1.2	0.6	3.6	2.4	3.5	2.4	2.0	1.3	2.1	1.2	1.3	0.6

Abbreviation definition: MED: Mean Euclidean distance; MMD: Mean Manhattan distance; SD: Standard deviation. C7: 7th cervical vertebra; PIIS: Posterior inferior iliac spine; TOC: Tip of coccyx. P1 – P6 correspond to the landmarks in Fig. 2.

Table 2: **Evaluation metrics on landmark prediction between the 6 predicted landmarks and their corresponding ground truth landmarks on the prospective test dataset.**

of C7 and TOC landmarks achieved the best performance (mean ± SD) compared with other landmarks in terms of MED and MMD (C7: MED = 1.0 ± 0.5, MMD = 1.2 ± 0.6; TOC: MED = 1.0 ± 0.5, MMD: 1.3 ± 0.6). The detection of left and right PIIS achieved an inferior performance (Left PIIS: MED = 1.6 ± 1.2, MMD = 2.0 ± 1.3; Right PIIS: MED = 1.7 ± 1.2, MMD = 2.1 ± 1.2). For the detection of left and right inferior scapular angles, average values of both MED and MMD achieved less than 4 pixels (Left inferior scapular angle: MED = 3.0 ± 2.1, MMD = 3.6 ± 2.4; Right inferior scapular angle: MED = 2.9 ± 2.3, MMD = 3.5 ± 2.4). For all 6 landmarks, values of both MED and MMD follow unimodal distribution quantitatively as analysed by violin plot, box plot, and scatter plot (Fig. 2a). Visual comparisons for landmark detection on back depth contours are presented in Fig. 2b with heatmaps of the 6 landmarks. We present the visual results of examples of landmark detections for 9 patients diagnosed with different AIS severities and types.

For landmark detection, the performance of 4 classical deep learning architectures with similar number of parameters was compared (Supplementary Table S2). For all 6 anatomical landmarks, HRNet[34] backbone achieves the best performance in terms of MED and MMD (Supplementary Table S3). The direct outputs of back landmark detection module are 6 landmark heatmaps. A high value means the probability of the landmark located at this corresponding position is high and vice versa. We used a probability threshold to filter the heatmaps to evaluate the quality of the output heatmaps. For different threshold values (from 0.1 to 0.6), HRNet achieved highest landmark retrieval rate for all 6 landmarks, and the retrieval rate was relatively stable when changing the threshold value compared with other architectures (Supplementary Table S4).

Performance of our deep learning framework for RCI synthesis

Patients with different severity levels and spinal curve types are presented to demonstrate the synthesis performance of our model on various cases (Fig. 3). The RGB image, depth image (in grayscale), synthesized RCIs and GT radiographs are displayed sequentially from the left to right. The normal-mild (Fig. 3a and b), moderate (Fig. 3c and d), and severe (Fig. 3e and f) AIS cases are displayed, and for each severity level (except Normal-mild), two cases with different curve types are visualized.

The pixel-wise metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) were unstable and inconsistent with the other metrics, and thus were not suitable to be used as measurements for RCI synthesis, after comparison of 12 combinations of generators and discriminators (Supplementary Table S5) in terms of 7 quantitative image quality metrics. Except PSNR and SSIM, the other 5 metrics, namely, 1) FID,[28] 2) VIF,[29] 3) LPIPS,[30] 4) NIQE,[31] and 5) BRISQUE,[32] performed consistently and indicated that a ResNet[35] with 9 blocks as generator and a 5 convolutional layer PatchGAN[36] as discriminator achieved the best performance.

Performance on Cobb angle prediction

The reliability of the synthesized RCIs for CA quantification had a strong correlation with the GT angles (R^2 = 0.984, p < 0.001) tested by linear regression. Additionally, the slope of the regression line was 45.99° comparable to the ideal value of 45°. According to the Bland–Altman plot (Fig. 4b), the mean difference of CAs obtained from GT and synthesized RCIs was minimal at −0.86°.

For each severity, three examples were visualized with different curve types to demonstrate the robustness of our model for heterogeneous cases (Fig. 5). Predicted CAs using synthesized RCIs were comparable with the angles measured using GT spine radiographs. The spine morphology of the synthesized RCIs was also close to the GT spine radiographs.

Performance on severity grading and curve type classification and prediction

For severity classifications (Table 3), both specificity (Sp) and accuracy (Acc) exhibited a relatively high score in grading all three severity levels. The negative predictive value (NPV) had a high score (0.981) for Normal-Mild cases (Table 3). The precision had the highest score

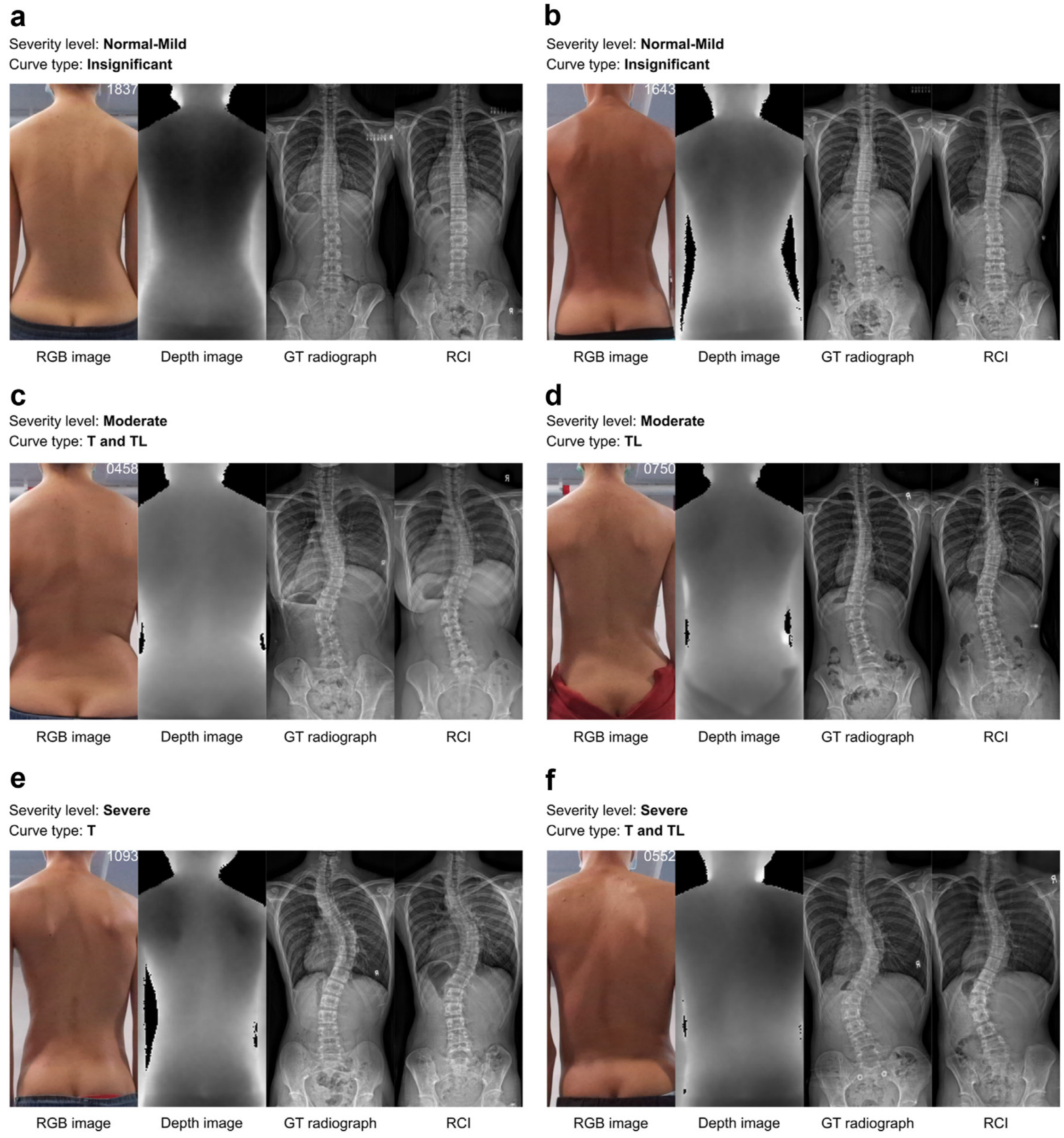

***Fig. 3:* RCI synthesis results from the RCI synthesis module**. Each subfigure presents a case with a certain severity level and curve type of scoliosis. From left to right, each subfigure presents the RGB image of the patient's nude back, depth image of the patient's nude back, the GT radiograph of the spine region and the synthesized RCI. **a and b** exhibit the RCI synthesis results for normal-mild cases. **c and d** show the moderate cases. Among them, the patient in. **c** has both T and TL curve while the one in. **d** has only TL curve. **e and f** present the severe cases. Among them, the patient in **e** has T curve while the patient in **f** has both T and TL curve. Abbreviation definition: RCI: Radiograph-comparable image; T: Thoracic; TL: Thoracolumbar; RGB: Red green blue; GT: Ground truth.

for both Moderate cases (0.962) and Severe cases (1.000) (Table 3). For curve type classification, the sensitivity (Se) and precision (Pr) achieved highest scores (T: Se = 0.978, Pr = 0.969; TL/L: Se = 0.974, Pr = 0.958). The confusion matrices for severity classification (Fig. 4c) comparison between synthesized RCI and GT spine radiographs, curve detection for T curve (Fig. 4d), curve detection for TL/L curve (Fig. 4e), and major curve (Fig. 4f) were illustrated.

Discussion

Medical image synthesis (MIS) has been fully studied to mutually transform inter- and intra-modality medical

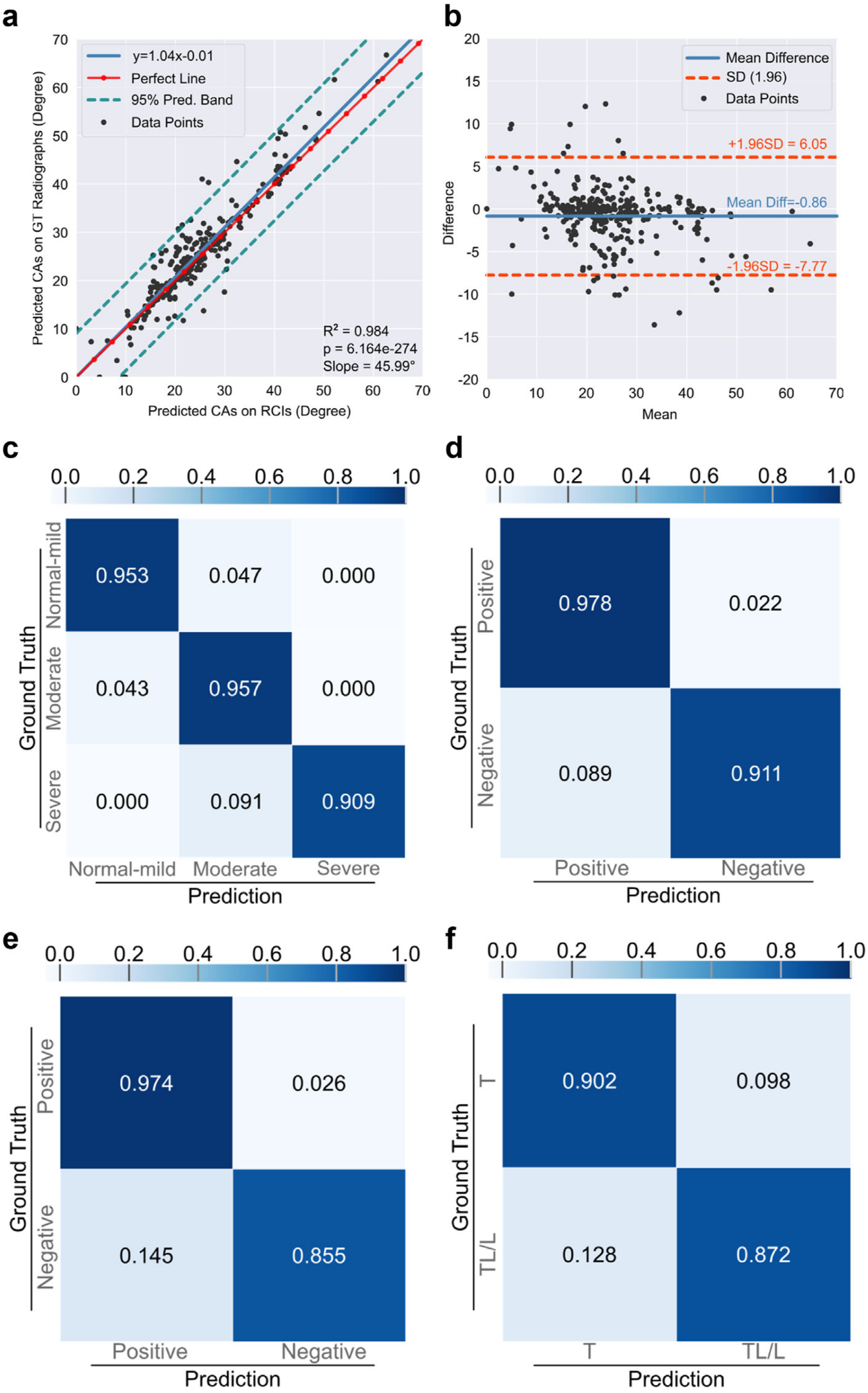

Fig. 4: **The performance evaluation of our model on CA prediction**. **a**, Linear regression analysis of CAs on the prospective test data. The x-axis denotes the CA predicted from our synthesized RCIs, while the y-axis denote the CA obtained from GT radiographs. **b**, Bland–Altman plots assessing the agreement of CAs obtained from synthesized and GT radiographs on the prospective test data. The x-axis represents the average degree of CAs obtained from synthesized and GT Radiographs, while y-axis denotes the difference between them. **c**, confusion matrix for the severity grading (3 types: normal-mild, moderate, and severe). **d–e**, confusion matrices for detection of the presence T and TL/L curves, respectively. **f**, confusion matrix for major curve type prediction. Abbreviation definition: CA: Cobb angle; RCI: Radiograph-comparable image;

images with deep learning. However, this study showed a few crucial and novel points for this topic. First, we demonstrate that an AI-powered system has the potential to synthesize RCI from optical images (RGBD images). There is potential to provide a radiation-free screening and analysis technique to assist AIS detection and diagnosis. Second, our experimental results show that the analytic results obtained from synthesized RCIs are coherent with the results from real radiographs (GT). In this regard, the developed light-based RCI synthesis system has the capacity to generate reliable RCI to substantively assist the clinical diagnosis procedure for AIS.

An accurate detection of the 6 anatomical landmarks with high confidence is crucial for RCI synthesis, since it provides useful information to assist the identification of spine morphology. With this in mind, the bareback landmark detection module was designed to output heatmaps of the 6 bareback landmarks, indicating both landmark position and probability of its presence. The validity and generalization of the bareback landmark detection module was examined on the prospective test dataset quantitatively and statistically, and we achieved less than 4 pixels error on average for each back landmark in terms of both MED and MMD measurements (Table 2). In addition, our model can accurately predict the landmarks from nude back images of patients with different severity levels of AIS (Fig. 3b).

To find the optimal deep learning architecture for back landmark detection, we compared 4 different classical frameworks (Supplementary Section 2.2). Considering the number of parameters can impact the capacity of the model,[37] for fairness, we adjusted the models accordingly to make them have roughly equivalent number of parameters by changing the number of filters, number of residual blocks, kernel and stride size of convolutional layers etc. Supplementary Table S2 provides a summary of the models. In addition to model parameters, we also compared the inference memory and floating-point operations (FLOPs) of different models. The metric inference memory refers to the required memory when using the model for testing while FLOPs is used to measure the model complexity. As shown, we controlled the number of parameters in different models to be similar to the capacity of different models. By minimizing the impacts of model capacity, model performance is then largely related to its architecture. Supplementary Table S3 compares the performance of different models for landmark detection. As shown, HRNet outperforms the other 3 deep learning models by a large margin, especially for the detection of left and right inferior scapular angle landmarks. Combining Supplementary Table S2 and Supplementary Table S3, HRNet uses a modest size of inference memory and relatively smaller computations (FLOPs) to achieve the best performance.

To evaluate the quality of the predicted landmark heatmaps, we used a threshold to filter the heatmap results. When increasing the threshold value, some landmark positions with low probability (smaller than the threshold) were filtered out and could not be retrieved. Supplementary Table S4 illustrates the retrieval rate under different probability thresholds for each of the 6 landmarks. As shown, compared with other deep learning architectures, HRNet performs consistently better on detection of all 6 landmarks with different probability thresholds. A landmark located at the position with a low probability value is usually not accurate. The probability threshold can help to filter out such inaccurate predictions. However, if the threshold is too high, some good predictions can be filtered out which will deteriorate the model performance. To decide a rational threshold, we plot the curve between the retrieval rate/MME/MED and probability threshold. As shown in Supplementary Fig. S7, we select 0.6 as the probability threshold to filter out bad predictions. With this threshold, our HRNet can still retrieve over 90% of the landmarks and the predicted landmarks with low probabilities can be removed as well.

The quality of synthesized RCIs was assessed in two aspects, namely the image quality and the usability of synthesized images in analytic clinical applications. Given the misalignment between RGBD images and radiographs, it was not appropriate to use pixel-wise image quality metrics (e.g., PSNR and SSIM), to evaluate the model performance. In this study, we introduced 5 metrics which measure the quality of the synthesized image in terms of the distribution or properties of extracted image features (FID, LPIPS and NIQE), image fidelity (VIF), and image spatial quality (BRISQUE). Supplementary Table S5 compares the performance of different metrics on synthesized RCIs. As shown, both PSNR and SSIM had low values, and the results were not consistent. In comparison, the measuring results obtained in terms of FID, LPIPS, NIQE, VIF, and BRISQUE were consistent, especially for the best two combinations of generator and discriminator.

The clinical quality and usability of synthesized RCIs were assessed in multiple analytic clinical applications, including the severity grading, curve type identification, and CA prediction. According to the data demographic

GT: Ground truth; T: Thoracic; TL/L: Thoracolumbar/lumbar; SD: Standard deviation. The p-value helps to determine the correlation between the predicted CA using real radiographs and RCIs, and the small value (p-value <0.0001) indicates they have strong correlation.

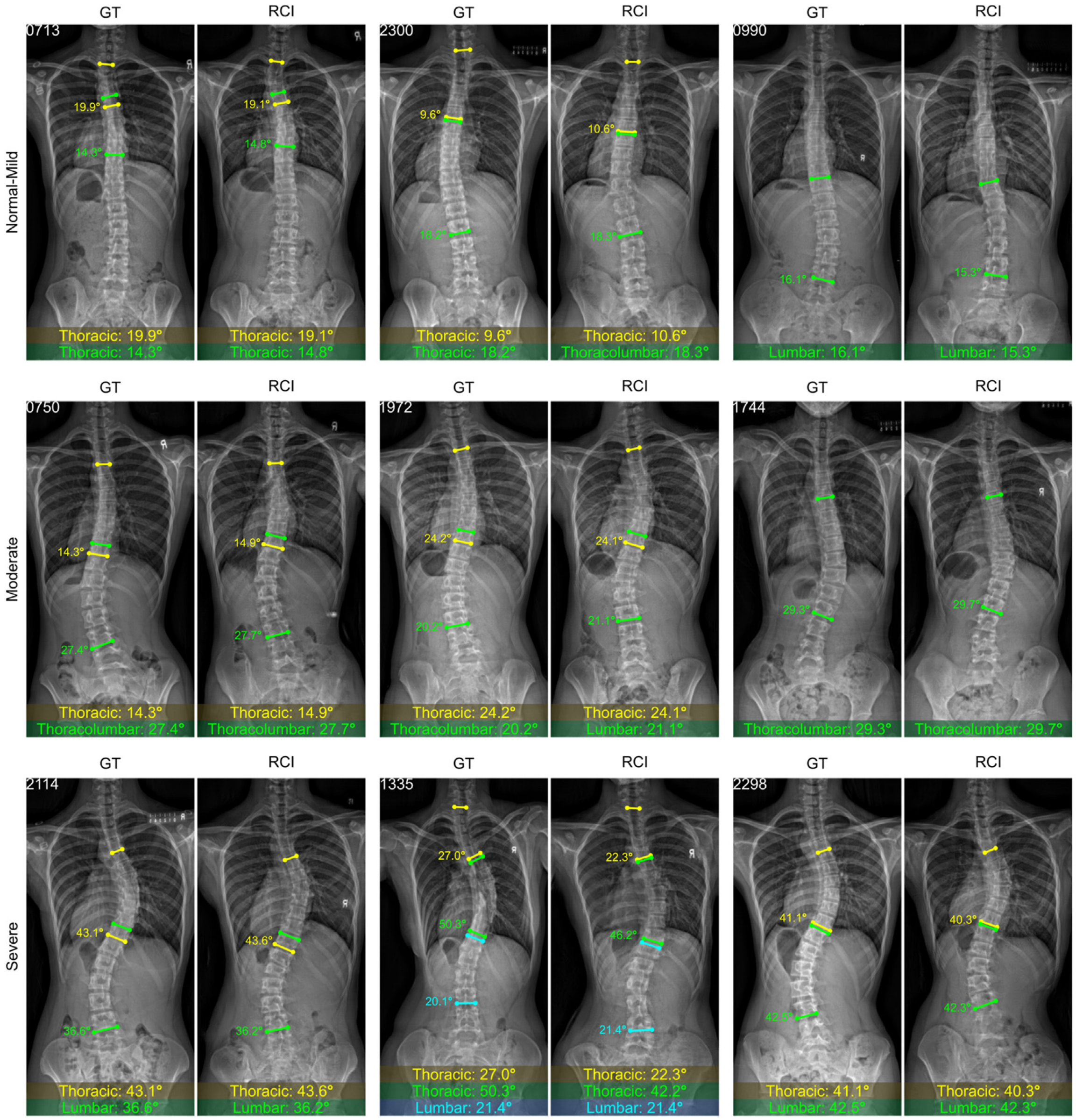

***Fig. 5:* Visual comparison of practical CA measuring in clinics using the original radiographs (left) and the synthesized RCI counterparts (right)**. The first row presents normal-mild cases with single or double curves. The second row presents moderate cases with single or double curves. The third row presents severe cases with double or triple curves. The patient number is printed on the top left corner on the GT radiograph to indicate different patients. The curve type and CA measuring results are printed on the bottom of each radiograph. Abbreviation definition: CA: Cobb angle; RCI: Radiograph-comparable image; GT: Ground truth.

in Fig. 1c, moderate patients take up about 60% of the participants. The second largest population is normal-mild (about 30%), and the severe group is the minority (about 10%). Correspondingly, from the synthesized RCIs, clinicians can distinguish the moderate and normal-mild patients with high sensitivity (>95%), while distinguishing the severe patients with relatively lower sensitivity (0.909), as shown in Table 3. Even so, the severity grading results are still satisfactory in terms of the high accuracy, and the performance on curve type classification can also validate the good quality of the synthesized RCIs.

The agreement of CA prediction results was examined using linear regression analysis, and the results indicated a strong correlation with the CAs predicted from GT radiographs (Fig. 4: R^2 = 0.984; p-value

Evaluation metrics	Severity level			Curve type	
	Normal-Mild	Moderate	Severe	T	TL/L
Sensitivity	0.953	0.957	0.909	0.978	0.974
Specificity	0.963	0.941	1.000	0.911	0.855
Precision	0.911	0.962	1.000	0.969	0.958
NPV	0.981	0.933	0.989	0.935	0.908
Accuracy	0.960	0.950	0.990	0.960	0.947

Abbreviation definition: NPV: Negative predictive value; T: Thoracic; TL/L: Thoracolumbar/Lumbar.

***Table 3*: Quantitative performance on severity grading and curve type identification on the prospective test dataset.**

<0.0001). In addition, the difference between predicted CAs from two spine alignments sources are close (mean difference = −0.86) (Fig. 4). In terms of visual results presented in Fig. 5, the alignment of spine is clearly synthesized, although there are mismatches in some vertebral regions, the CA can be accurately predicted.

Our study has a few limitations. The performance of the CA prediction largely depends on the quality of the synthesized RCIs which further relies on the accuracy of detected anatomical landmarks. Considering the nude back features of obese patients may not be obvious enough to obtain accurate landmarks, no overweight patients are recruited in this study. Another potential limitation is skin colour. Since most of the participants are Southeast Asian Chinese, the images in the dataset used for training model may not be able to accurately represent the skin colour of other population groups. In addition, the system was tested in two centres following the same procedure to collect data with the same clinical assessment standard. The performance of our model may reduce when directly applied to another centre. We are planning an international multi-centre trial to further assess the reliability of our system and device. To investigate the impacts of sample size, the curve-fitting method has been used. However, such a method is originally proposed to study how sample size affects classification model. Therefore, the investigation results may be not accurate reflect the impacts of sample size to our models. Nevertheless, some analysis regarding overall trends still has certain reference value. For example, when the sample size of training data decreases, the stability and performance of both models deteriorate. Furthermore, when the training sample size is very small, both models have a high risk of overfitting.

In summary, we deployed the first prospectively tested auto-alignment analysis model for spine malalignment analysis using deep learning and RGBD technologies with no radiation. With future multi-centre validation, our platform can better assist clinicians and clinical research in large volumes.

Contributors

NM implemented the deep learning models, evaluated the model performance, conducted statistical analysis, accessed and verified the underlying data. KKW co-supervised the pipeline design and implementation. MXZ collected and managed the clinical data and conducted the demographic analysis. JPC designed the study, supervised the clinical parameters and significance, and was responsible for the annotations and labelling. TZ designed the study, searched the literature, supervised the data collection and model implementation, accessed and verified the underlying data, developed the RGBD imaging device, and was responsible for result assessments. NM and TZ drafted the manuscript, while all authors reviewed the manuscript for important intellectual content and approved the final manuscript.

Data sharing statement

Restrictions apply to the availability of the training and validation datasets, which were used with permission of the participants of this study. Anonymized prospective testing data may be available for research purposes from the corresponding author on a reasonable request and under the related institution policy. The deep learning models in our light-based RCI synthesis system were developed and deployed using standard model libraries and the PyTorch framework. The original HRNet, ResNet and PatchGAN models are publicly available. Custom version of these models is specific to our development system and tasks. The codes may be available for research purposes from the corresponding author on a reasonable request and under the related institution policy.

Editor note

The Lancet Group takes a neutral position with respect to territorial claims in published maps and institutional affiliations.

Declaration of interests

All authors declare there is no conflict of interest.

Acknowledgments

We sincerely appreciated the support from the Innovation and Technology Fund (MRP/038/20X), and Health Services Research Fund (HMRF) 08192266.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102050.

References

1 Cobb J. Outline for the study of scoliosis. *Instructional Course Lectures AAOS*. 1948;5:261–275.

2 Fong DYT, Cheung KMC, Wong YW, et al. A population-based cohort study of 394,401 children followed for 10 years exhibits sustained effectiveness of scoliosis screening. *Spine J*. 2015;15(5):825–833.

3 Luk KD, Lee CF, Cheung KM, et al. Clinical effectiveness of school screening for adolescent idiopathic scoliosis: a large population-based retrospective cohort study. *Spine*. 2010;35(17):1607–1614.

4 Weinstein SL, Dolan LA, Spratt KF, Peterson KK, Spoonamore MJ, Ponseti IV. Health and function of patients with untreated idiopathic scoliosis: a 50-year natural history study. *JAMA*. 2003; 289(5):559–567.

5 Yang J, Zhang K, Fan H, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. *Commun Biol*. 2019;2(1):1–8.

6 Cheung JPY, Zhang T, Bow C, Kwan K, Sze KY, Cheung KMC. The crooked rod sign: a new radiological sign to detect deformed threads in the distraction mechanism of magnetically controlled growing rods and a mode of distraction failure. *Spine*. 2020;45(6):E346–E351.

7 Côté P, Kreitz BG, Cassidy JD, Dzus AK, Martel J. A study of the diagnostic accuracy and reliability of the Scoliometer and Adam's forward bend test. *Spine*. 1998;23(7):796–802.

8 Fong DYT, Lee CF, Cheung KMC, et al. A meta-analysis of the clinical effectiveness of school scoliosis screening. *Spine*. 2010;35(10):1061–1071.

9 Chung N, Cheng YH, Po HL, et al. Spinal phantom comparability study of Cobb angle measurement of scoliosis using digital radiographic imaging. *J Orthop Translat*. 2018;15:81–90.

10 Grauers A, Einarsdottir E, Gerdhem P. Genetics and pathogenesis of idiopathic scoliosis. *Scoliosis Spinal Disord*. 2016;11(1):1–7.

11 Rodemann HP, Blaese MA. Responses of normal cells to ionizing radiation. Elsevier *Semin Radiat Oncol*. 2007;17(2):81–88.

12 Brody AS, Frush DP, Huda W, Brent RL. Radiation risk to children from computed tomography. *Pediatrics*. 2007;120(3):677–682.

13 Cheung CWJ, Law SY, Zheng YP. *Development of 3-D ultrasound system for assessment of adolescent idiopathic scoliosis (AIS): and system validation. International Conference of the IEEE Engineering in Medicine and Biology Society; 2013*. IEEE; 2013:6474–6477.

14 Czaprowski D, Pawłowska P, Gębicka A, Sitarski D, Kotwicki T. Intra-and interobserver repeatability of the assessment of anteroposterior curvatures of the spine using Saunders digital inclinometer. *Ortop Traumatol Rehabil*. 2012;14(2):145–153.

15 Melvin M, Sylvia M, Udo W, Helmut S, Paletta JR, Adrian S. Reproducibility of rasterstereography for kyphotic and lordotic angles, trunk length, and trunk inclination: a reliability study. *Spine*. 2010;35(14):1353–1358.

16 Perriman DM, Scarvell JM, Hughes AR, Ashman B, Lueck CJ, Smith PN. Validation of the flexible electrogoniometer for measuring thoracic kyphosis. *Spine*. 2010;35(14):E633–E640.

17 Meng N, Ge Z, Zeng T, Lam EYN. A deep generative model for light field reconstruction. *IEEE Access*. 2020;8:116052–116063.

18 Meng N, Li K, Liu J, Lam EY. Light field view synthesis via aperture disparity and warping confidence map. *IEEE Trans Image Process*. 2021;30:3908–3921.

19 Freitag MT, Fenchel M, Bäumer P, et al. Improved clinical workflow for simultaneous whole-body PET/MRI using high-resolution CAIPIRINHA-accelerated MR-based attenuation correction. *Eur J Radiol*. 2017;96:12–20.

20 Zhang T, Li Y, Cheung JPY, Dokos S, Wong KYK. Learning-based coronal spine alignment prediction using smartphone-acquired scoliosis radiograph images. *IEEE Access*. 2021;9:38287–38295.

21 Devic S. MRI simulation for radiotherapy treatment planning. *Med Phys*. 2012;39(11):6701–6711.

22 Liu F, Jang H, Kijowski R, Bradshaw T, McMillan AB. Deep learning MR imaging–based attenuation correction for PET/MR imaging. *Radiology*. 2018;286(2):676–684.

23 Dong X, Lei Y, Tian S, et al. Synthetic MRI-aided multi-organ segmentation on male pelvic CT using cycle consistent deep attention network. *Radiother Oncol*. 2019;141:192–199.

24 Yang Q, Yan P, Zhang Y, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. 2018;37(6):1348–1357.

25 Microsoft azure kinect DK. https://azure.microsoft.com/en-us/services/kinect-dk/; 2020. Accessed June 16, 2018.

26 Mak T, Cheung PWH, Zhang T, Cheung JPY. Patterns of coronal and sagittal deformities in adolescent idiopathic scoliosis. *BMC Musculoskelet Disord*. 2021;22(1):1–10.

27 Meng N, Cheung JP, Wong KYK, et al. An artificial intelligence powered platform for auto-analyses of spine alignment irrespective of image quality with prospective validation. *eClinicalMedicine*. 2022;43:101252.

28 Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv Neural Inf Process Syst*. 2017;30.

29 Sheikh HR, Bovik AC. Image information and visual quality. *IEEE Trans Image Process*. 2006;15(2):430–444.

30 Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. *IEEE Conf Comput Vis Pattern Recogn*. 2018;2018:586–595.

31 Mittal A, Soundararajan R, Bovik AC. Making a "completely blind" image quality analyzer. *IEEE Signal Process Lett*. 2012;20(3):209–212.

32 Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain. *IEEE Trans Image Process*. 2012;21(12):4695–4708.

33 Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can Assoc Radiol J*. 2019;70(4):344–353.

34 Wang J, Sun K, Cheng T, et al. *Deep high-resolution representation learning for visual recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence; 2020.

35 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *IEEE Conf Comput Vis Pattern Recogn*. 2016;2016:770–778.

36 Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. *IEEE Conf Comput Vis Pattern Recogn*. 2017;2017:1125–1134.

37 Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. *Commun ACM*. 2021;64(3):107–115.