Automated measurement of leg length discrepancy from infancy to adolescence based on cascaded LLDNet and comprehensive assessment

Qiang Zheng[1]^, Bin Liu[1], Xiangrong Tong[1], Jungang Liu[2], Jian Wang[2], Lin Zhang[2]^

[1]School of Computer and Control Engineering, Yantai University, Yantai, China; [2]Department of Radiology, Xiamen Children's Hospital, Children's Hospital of Fudan University at Xiamen, Xiamen, China

Contributions: (I) Conception and design: Q Zheng, L Zhang; (II) Administrative support: Q Zheng, L Zhang; (III) Provision of study materials or patients: L Zhang, J Liu, J Wang; (IV) Collection and assembly of data: Q Zheng, L Zhang, J Liu, J Wang, B Liu; (V) Data analysis and interpretation: Q Zheng, B Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Lin Zhang, MD. Associate Chief Physician, Department of Radiology, Xiamen Children's Hospital, Children's Hospital of Fudan University at Xiamen, No. 92, Yibin Road, Huli District, Xiamen 361006, China. Email: lzhang_XMChospital@hotmail.com.

Background: Deep learning (DL) has been suggested for the automated measurement of leg length discrepancy (LLD) on radiographs, which could free up time for pediatric radiologists to focus on value-adding duties. The purpose of our study was to develop a unified solution using DL for both automated LLD measurements and comprehensive assessments in a large and comprehensive radiographic dataset covering children at all stages, from infancy to adolescence, and with a wide range of diagnoses.

Methods: The bilateral femurs and tibias were segmented by a cascaded convolutional neural network (CNN), referred to as LLDNet. Each LLDNet was conducted through use of residual blocks to learn more abundant features, a residual convolutional block attention module (Res-CBAM) to integrate both spatial and channel attention mechanisms, and an attention gate structure to alleviate the semantic gap. The leg length was calculated by localizing anatomical landmarks and computing the distances between them. A comprehensive assessment based on 9 indices (5 similarity indices and 4 stability indices) and the paired Wilcoxon signed-rank test was undertaken to demonstrate the superiority of the cascaded LLDNet for segmenting pediatric legs through comparison with alternative DL models, including ResUNet, TransUNet, and the single LLDNet. Furthermore, the consistency between the ground truth and the DL-calculated measurements of leg length was also comprehensively evaluated, based on 5 indices and a Bland-Altman analysis. The sensitivity and specificity of LLD >5 mm were also calculated.

Results: A total of 976 children were identified (0–19 years old; male/female 522/454; 520 children between 0 and 2 years, 456 children older than 2 years, 4 children excluded). Experiments demonstrated that the proposed cascaded LLDNet achieved the best pediatric leg segmentation in both similarity indices (0.5–1% increase; P<0.05) and stability indices (13–47% percentage decrease; P<0.05) compared with the alternative DL methods. A high consistency of LLD measurements between DL and the ground truth was also observed using Bland-Altman analysis [Pearson correlation coefficient (PCC) =0.94; mean bias =0.003 cm]. The sensitivity and specificity established for LLD >5 mm were 0.792 and 0.962, respectively, while those for LLD >10 mm were 0.938 and 0.992, respectively.

Conclusions: The cascaded LLDNet was able to achieve promising pediatric leg segmentation and LLD measurement on radiography. A comprehensive assessment in terms of similarity, stability, and measurement consistency is essential in computer-aided LLD measurement of pediatric patients.

Keywords: Leg length discrepancy (LLD); deep learning (DL); radiograph

---

^ ORCID: Qiang Zheng, 0000-0002-7853-8033; Lin Zhang, 0000-0003-3607-2121.

Introduction

Leg length discrepancy (LLD) is an orthopedic problem that frequently causes musculoskeletal disorders in children, such as gait deviations, scoliosis, low back pain, osteoarthritis, and postural control (1). Particularly, children with an LLD greater than 5 mm could have an increased risk of hip, knee, or back problems (2,3). Therefore, accurate and reliable LLD measurement is crucial for planning appropriate treatment (4-6).

Radiography of bilateral lower limbs is considered a standard approach to measuring LLD (1,7,8). Pediatric radiologists used to manually measure the LLD on radiographs from the upper edge of the femoral head to the distal tibia; however, while cognitively simple, this task is labor-intensive. A previous study (9) demonstrated that radiology technologists can be rapidly trained to measure LLD as precisely as can a board-certified pediatric radiologist. Therefore, the delegation of this time-consuming task to artificial intelligence (AI)-powered assistants is becoming increasingly important.

Deep learning (DL) is a powerful tool and, due to its successful application in a range of settings, it is expected that DL will be able to perform the simple but labor-intensive measurement component of radiologic examinations (10,11). In prior studies, DL was adopted for the automated and rapid measurement of LLD, including leg segmentation-based measurement (12) and anatomical landmark localization-based measurement (13). In these studies, both approaches employed convolutional neural networks (CNNs) from different perspectives and achieved promising accuracy. However, these studies were only performed on children from a preschool stage and of a relatively small cohort. Children with LLD, particularly, infants and toddlers, pose challenges in completing clinical routines.

Children will experience many remarkable changes in terms of both the structure and alignment of legs during their development (14). For instance, approximately 90% of cases of capital femoral epiphysis can be observed on radiographs as early as 200 days after birth (15). Sugawara *et al.* (16) and Garn *et al.* (17) reported that the 50th percentile of age at the appearance of the femoral head was 6 months in children with normal flexion and 8 months in children who had developmental dysplasia of the hip. Ossification of the trochanteric apophysis begins at approximately 4 years of age in both girls and boys, but that of the distal femoral epiphysis can occur prior to birth (18). Furthermore, the knee joint space of infants is demonstrated on radiographs to be wider than that observed in older children on radiographs due to the relatively small ossification center of the distal femoral epiphysis and proximal tibial epiphysis, as well as the surrounding epiphyseal cartilage, which is not visible on X-ray plain film. *Figure 1* demonstrates a series of radiographs covering children at different stages.

Upon the success of a previous study (12), the present study adopted the leg segmentation-based strategy for automated LLD measurement. First, a cascaded CNN (referred to as "cascaded LLDNet") was devised for the accurate pediatric leg segmentation. Second, a comprehensive assessment was performed; however, the limited spatial overlap was insufficient for evaluating the improper segmentation for LLD measurement. Our hypothesis was that the cascaded LLDNet could achieve promising pediatric leg segmentation and LLD measurement based on a comprehensive assessment of a large and comprehensive radiographic dataset comprising children at all stages, from infancy to adolescence, and with a wide range of diagnoses. We present the following article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-22-282/rc).

Methods

Study participants

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by institutional ethics board of Xiamen Children's Hospital, and individual consent for this retrospective analysis was waived. A total of 976 children who underwent full-length radiography of their lower limbs from the hips to the ankles filmed under anteroposterior single exposure were identified between January 2020 and December 2020 from the database of a local children's hospital. The inclusion criteria were as follows: (I) adequate technical quality of radiograph; and (II) imaging field of view (FOV) that covered the entire bilateral femurs and tibias. The
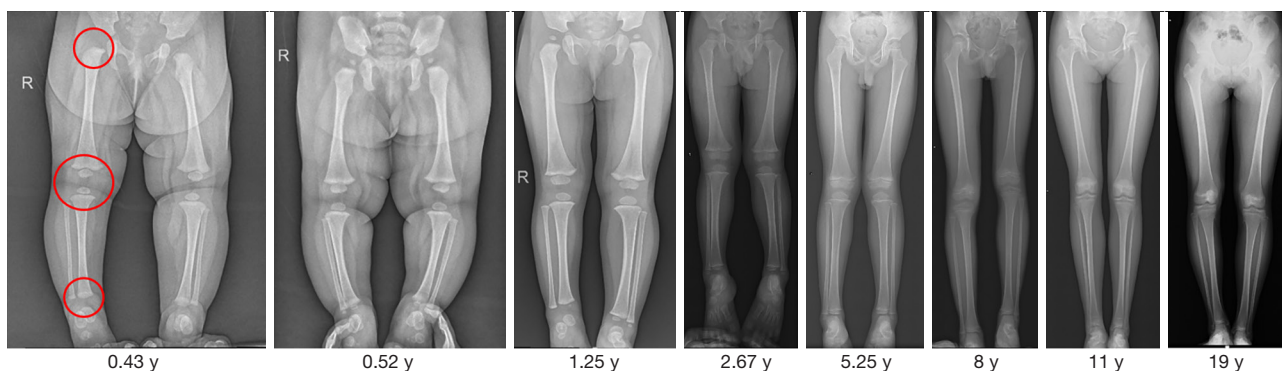
Figure 1 Radiographs covering children at all stages. The red circles show regions including capital femoral epiphysis, the distal femoral epiphysis, the proximal and distal tibial epiphysis, which will experience remarkable changes in terms of structure and alignment of legs during childhood development. R, right; y, year.

exclusion criteria were as follows: imaging FOV that did not include the entire bilateral femurs and tibias. We had 3 patients with internal fixation devices, and those patients were excluded. One radiograph did not include the distal tibias bilaterally in the imaging FOV, and thus it was excluded. Ultimately, there were 4 radiographs excluded based on the exclusion criteria (4/976, 0.41%), and 972 children were included in our study (520 children between 0–2 years and 456 children older than 2 years). The radiographs with leg deformities or skeletal dysplasias were not excluded because we wanted to establish a dataset with a broad spectrum of clinically common diseases.

The full-length radiography was acquired by 2 imaging acquisition systems, the DigitalDiagnost 3 digital radiography (DR) scanning unit (Philips Healthcare, Andover, MA, USA) and the SONIALVISION Plus multipurpose digital radiography/fluoroscopy (R/F) system (Shimadzu Corporation, Kyoto, Japan). For children under 2 years, we used the DR scanning unit because the FOV was sufficiently large to scan the entirety of the lower limbs in a single exposure. For older children over 2 years, we used the multipurpose R/F system perform full-length radiography of both lower limbs. This system determined the acquisition start point and end point under fluoroscopy, which was followed by acquisition of the images and then stitching of the images to generate full-length radiographs with a postprocessing workstation.

In our study, some patients had no clinical diseases but were observed to have an abnormal appearance or trauma in both lower limbs [referred to as "children with no clinical disease (NCD) group"]. The other patients in

the radiographic dataset carried a wide range of diagnoses [referred as "children with various clinical diseases (VCD) group"], including orthopedic hardware, Beckwith-Wiedemann syndrome, bow legs, cerebral palsy, developmental dysplasia of the hip, fibrous dysplasia, fibrous cortical defect, foot tumor, growth and intellectual disability, knock knees, Langerhans cell histiocytosis, neurofibromatosis type 1, osteochondroma, scoliosis, Shwachman-Diamond syndrome, hip synovitis, tethered cord syndrome, and venous malformation.

The demographic information is summarized in *Table 1*. There were no significant differences between the NCD and VCD groups regarding femoral length, tibial length, or full leg length (P>0.05, *t*-test), but the LLD was significantly different between the NCD and VCD groups (P<0.05, *t*-test). This demonstrated that clinical disease in the VCD group was associated with pediatric LLD.

Ground truth

Since the present study adopted the leg segmentation–based strategy for automatic LLD measurement, the ground truth consisting of manual segmentation served as the standard masks and leg length measurement to validate the accuracy of the computer-aided approach.

Manual segmentation was performed by a senior attending pediatric radiologist with more than 5 years' working experience using ITK-SNAP v3.8.0 software (http://www.itksnap.org/pmwiki/pmwiki.php). The leg length was measured according to the anatomical landmarks adopted by Zheng *et al.* (12).

Table 1 Demographic information

Age	Groups	Number (N=972)	Age (years)	Sex (M/F)	Femoral length (cm)	Tibial length (cm)	Full leg length (cm)	Full LLD (cm)
0–2 years	NCD group	292	1.39±0.37	143/149	18.2±1.9 (7.2–21.8)	16.0±1.6 (8.2–19.3)	34.2±3.5 (15.9–40.5)	0.17±0.15 (0–0.83)
	VCD group	226	1.34±0.39	120/106	17.9±1.9 (7.8–23.8)	16.0±1.6 (7.9–20.9)	34.0±3.6 (15.8–44.7)	0.29±0.35 (0–2.4)
>2 years	NCD group	199	5.20±2.46	100/99	26.9±5.8 (18.5–47.4)	22.7±4.4 (15.8–37.6)	49.7±10.3 (34.3–84.7)	0.25±0.19 (0–0.9)
	VCD group	255	5.01±2.81	157/98	26.0±6.0 (15.6–45.6)	21.9±4.5 (13.9–38.0)	48.0±10.5 (29.5–83.7)	0.47±0.73 (0–4.4)

The NCD group includes children with no clinical disease, while the VCD group includes children with various clinical disease. No significant difference was observed in age, sex, femoral length, tibial length, or full leg length between the NCD and VCD groups, but a significant difference was observed in full LLD (chi-squared test for sex and *t*-test for others). Data are shown as mean ± SD and (ranges). NCD, children with no clinical disease; VCD, children with various clinical diseases; M, male; F, female; LLD, leg length discrepancy.
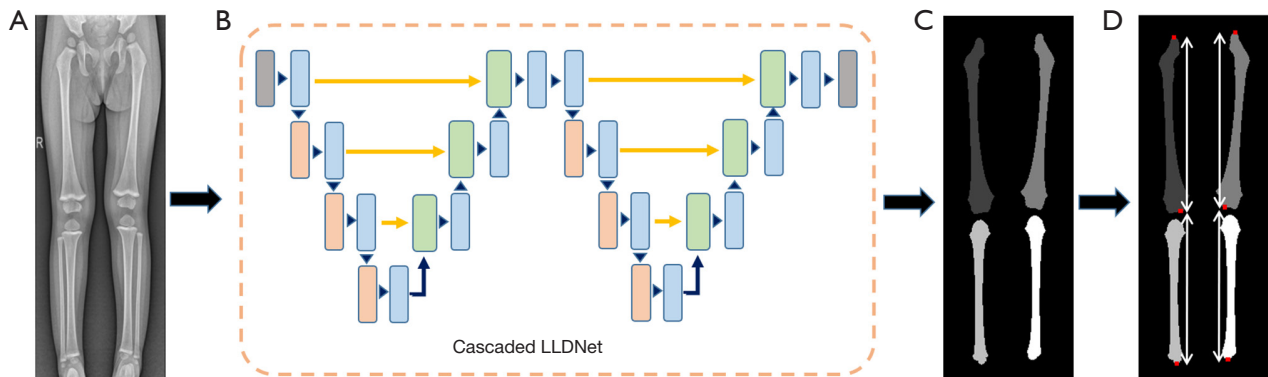
Figure 2 Schematic diagram of the cascaded LLDNet to accomplish simultaneous bilateral segmentation of femurs and tibias. (A) Image of area to be segmented. (B) The cascaded LLDNet model. (C) Segmentation of femurs and tibias. (D) Anatomical landmarks and leg length measurement. LLD, leg length discrepancy.

DL-based measurement of LLD

The proposed leg segmentation-based strategy typically comprised 4 stages: segmentation of pediatric legs by a cascaded LLDNet, postprocessing of segmentation masks, identification of anatomical landmarks, and measurements of the distances between them.

In the first stage of pediatric leg segmentation, we established a cascaded LLDNet to perform bilateral segmentation of the femurs and tibias in each radiograph (*Figure 2*). Specifically, the CNN model cascaded 2 identical, U-shaped network structures (each referred to as an LLDNet), with the first LLDNet being able to generate a probability matrix for the bilateral femur and tibia, and the second LLDNet being fed the probability matrix to segment the bilateral femurs and tibia in each radiograph.

Each cascaded LLDNet was conducted under the popular framework of U-Net but was improved via adoption of residual blocks (*Figure 3A*). The benefits of employing residual blocks included learning more features and accelerating convergence. The first residual block was followed by a residual convolutional block attention module (Res-CBAM), as shown in *Figure 3B*. The Res-CBAM consisted of a convolutional layer, CBAM, batch normalization, and rectified linear unit, in which the CBAM integrated both spatial and channel attention mechanisms to enhance discriminating feature maps (19). Additionally, the attention gate structure, as seen in *Figure 3C*, was added to skip connections to alleviate the semantic gap and highlight the target area in segmentation.
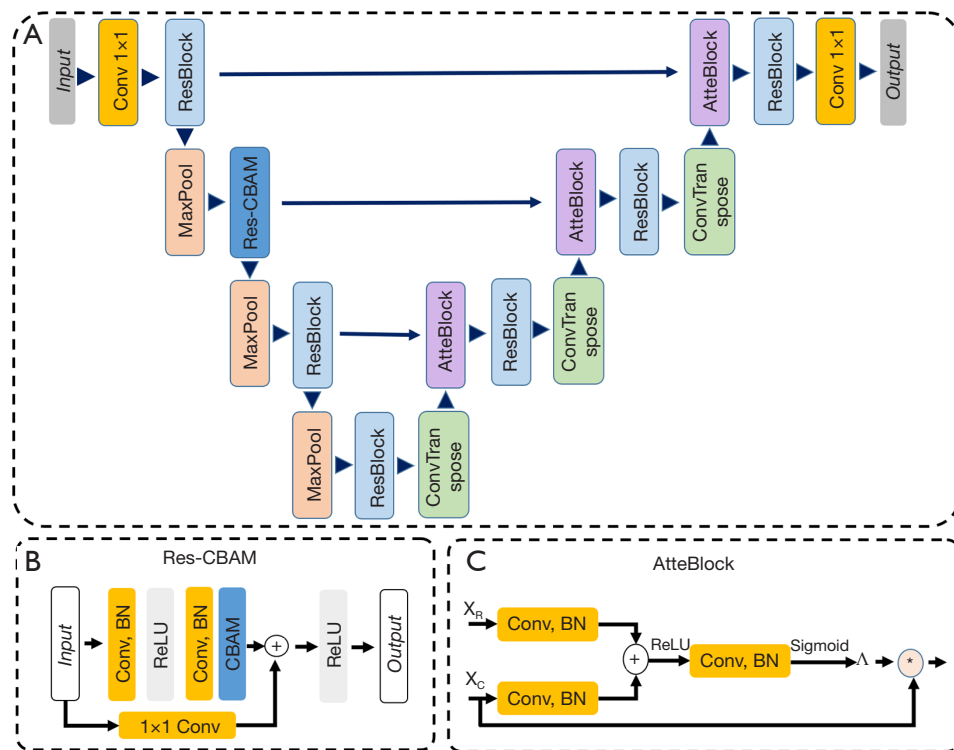
Figure 3 The LLDNet network. (A) The structure of the LLDNet network. Both the Maxpool and ConvTranspose were performed with the kernels of 2×2 and a stride of 2. (B) The Res-CBAM consisted of a convolutional layer, CBAM, BN, and ReLU. (C) AtteBlock: X_R indicates the feature map generated by ResBlock, X_C indicates the feature map generated by the ConvTranspose, Λ indicates the weight matrix, and * indicates voxel-wise multiplication for the feature maps. LLD, leg length discrepancy; Res-CBAM, residual convolutional block attention module; BN, batch normalization; ReLU, rectified linear unit.

To assess the generalization of the automatic measurement of leg length difference, a 5-fold cross-validation was adopted in our study. Specifically, the participants were split into a training set (60% of the data, 3 groups), a validation set (20% of the data, 1 group), and a testing set (20% of the data, 1 group) with even distribution of age and presence or absence of clinical disease. The procedure was performed 5 times to ensure each participant could be tested.

PyTorch deployed on the Nvidia GeForce Ray Tracing Texel eXtreme (RTX) 2080 graphics processing unit (GPU; NVIDIA Corp, Santa Clara, CA, USA) was used to train the CNN model in an end-to-end mode with the input images resized to 448×160 pixels, scaled to [0, 1], and normalized with a max–min standardized method. The Dice loss function $DiceLoss=1-2|A\cap B|/(|A|+|B|)$ was employed, where A and B were manual and automatic segmentations, and $|\cdot|$ was the volume of the corresponding segmentation mask. The batch size was 5, and the training time was

100 epochs. The initial learning rate was 1×10^{-4} and decayed by 10% every 30 epochs.

In the second stage of the postprocessing of segmentation masks, a maximum connected component analysis was performed to detect the pediatric legs in the segmentation label images. Specifically, given the 4 segmentation labels of left femur, left tibia, right femur, and right tibia in an image (referred to as a "4-label image"), we generated 4 new separate images, in which each image only contained on1e label (referred to as a "1-label image"). Then, the maximum connected region was recognized and reserved in each 1-label image, after which it was remerged into a 4-label image, which was treated as the final pediatric leg segmentation map.

In the third stage of the identification of anatomical landmarks, the same 3 landmarks used in a previous study (12) were determined for each limb: the apex of the femoral head (AFH; defined as the uppermost point of the femoral head, and if the femoral head of an infant less than 6 months
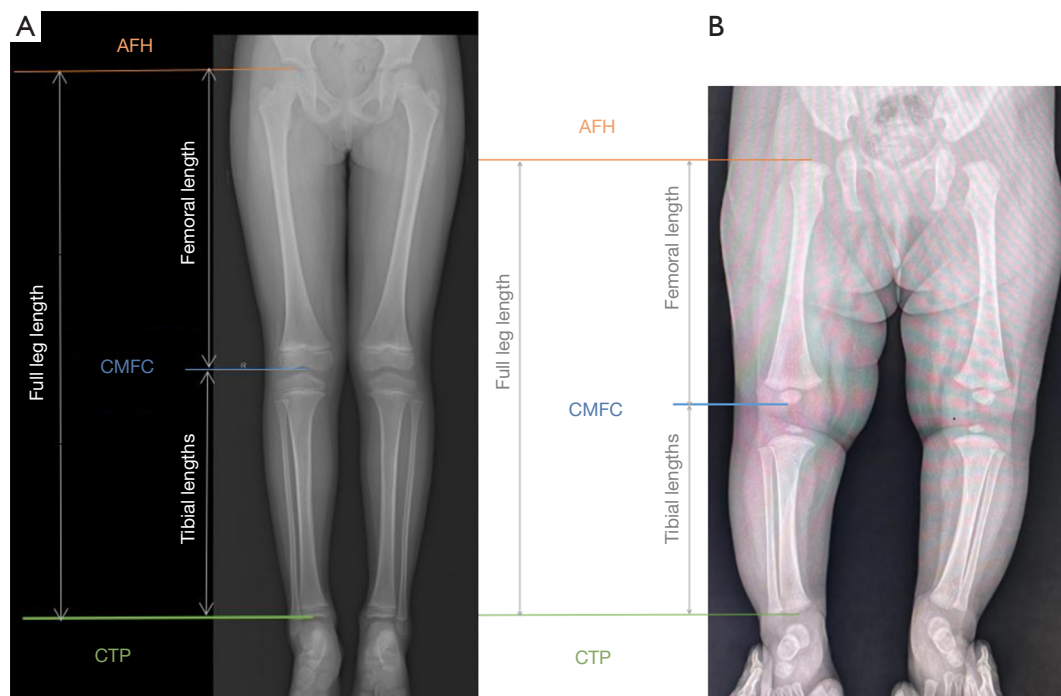
Figure 4 Full-length anteroposterior radiographs of 2 children. (A) Full-length anteroposterior radiograph of a 4-year-old girl. The femoral length is defined as the vertical distance from the AFH to the CMFC. The tibial length is defined as the vertical distance from the CMFC to the CTP. The full leg length is defined as the sum of the femoral and tibial lengths. LLD is defined as the length difference between the right and left limbs. (B) Full-length anteroposterior radiograph of a 3-month-old girl. Different with (A), the femoral heads were not ossified. The femoral length is defined as the vertical distance from the ossified upper femoral apex instead of the AFH to the CMFC. The tibial length is defined as the vertical distance from the CMFC to the CTP. AFH, apex of the femoral head; CMFC, convexity of the medial femoral condyle; CTP, center of the tibial plafond; LLD, leg length discrepancy.

old was not ossified completely, the ossified upper femoral apex was used to replace the AFH for measurement), the convexity of the medial femoral condyle (CMFC; defined as the bony protrusion on the inside edge of the bottom of the femur bone), and the center of the tibial plafond (CTP; defined as the center of the distal end of the tibia). The full leg length was defined as the sum of the femoral and tibial bone lengths, and the LLD was defined as the difference between the right and left full leg lengths. The femoral length was defined as the vertical distance between the AFH and the CMFC, and the tibial length was defined as the vertical distance between the CMFC and the CTP (*Figure 4*).

In the fourth stage of the measurement of the distance between landmarks, the length of the femurs, tibias, and full legs were calculated between corresponding anatomical landmarks (*Figure 2D*). Specifically, the femoral and tibial lengths were defined as the vertical distances between the AFH and CMFC and between the CMFC

and TP, respectively. The full limb length was the sum of the femoral and tibial lengths. LLD was defined as the difference in the length between the right and left limbs.

Comprehensive assessment and statistical analysis

The comprehensive assessment in this study included the evaluation of the following: (I) the accuracy of the segmentation of pediatric legs based on 9 indices and (II) the consistency of the measurement of leg length between DL and ground truth based on 5 indices. Using the comprehensive indices above, we compared the proposed cascaded LLDNet with alternative DL models, including ResUNet (12), TransUNet (20), and the single LLDNet to demonstrate the superiority of the cascaded LLDNet.

To assess the accuracy of the segmentation of the pediatric leg, 9 indices were employed (21), including Dice, Jaccard, precision, recall, mean distance (MD), Harsdorff

858

Zheng et al. Cascaded LLDNet for pediatric LLD measurement

distance (HD), HD95, average symmetric surface distance (ASSD), and root mean square deviation (RMSD). The above metrics comprehensively measured the segmentation in terms of similarity (Dice, Jaccard, precision, and recall) and stability (MD, HD, HD95, ASSD, and RMSD). Additionally, the paired Wilcoxon signed-rank test (MATLAB R2018b; MathWorks, Natick, MA, USA) was also employed to statistically compare the accuracy of the segmentation between different DL models. A value of P<0.05 indicated a significant improvement of the proposed cascaded LLDNet over alternative DL models in terms of the 9 indices. Given manual segmentation A and automatic segmentation B, the above 9 indices were defined as follows:

$$Dice = 2\frac{V(A \cap B)}{V(A) + V(B)}, \ Jaccard = \frac{V(A \cap B)}{V(A \cup B)} \quad [1]$$

$$Precision = \frac{V(A \cap B)}{V(B)}, \ Recall = \frac{V(A \cap B)}{V(A)} \quad [2]$$

$$MD = mean_{e \in \partial A}\left(min_{f \in \partial B} d(e, f)\right) \quad [3]$$

$$HD = \max\left(H(A, B), H(B, A)\right),$$
$$\text{where } H(A, B) = max_{e \in \partial A}\left(min_{f \in \partial B} d(e, f)\right) \quad [4]$$

HD95 is similar to HD, expect that 5% data points with the largest distance are removed before the calculation, as follows:

$$ASSD = \frac{mean_{e \in \partial A}\left(min_{f \in \partial B} d(e, f)\right) + mean_{e \in \partial B}\left(min_{f \in \partial A} d(e, f)\right)}{2} \quad [5]$$

$$RMSD = \frac{\sqrt{D_A^2 + D_B^2}}{card\{\partial A\} + card\{\partial B\}}, \text{ where } D_A^2 = \sum_{e \in \partial A}\left(min_{f \in \partial B} d(e, f)\right) \quad [6]$$

To assess the consistency of the measurement of leg length between DL and ground truth, 5 indices, including the Pearson correlation coefficient (PCC), mean squared error (MSE), mean absolute error (MAE), maximum absolute error (MaxAE), and concordance correlation coefficient (CCC), were employed. Given the ground truth y and the predicted measurement $\hat{y}$ by DL, the MSE and MAE were defined as follows:

$$Mean \ Squared \ Error = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2 \quad [7]$$

$$Mean \ Absolute \ Error = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|^2 \quad [8]$$

Additionally, the Bland-Altman plot was adopted to visualize the differences in the measurements of leg length between the manual and automatic measurements. The mean bias and the coefficient of repeatability (RPC) from the Bland-Altman analysis were calculated. Given the 2 measurements, y_2 and y_1, the RPC was defined as follows:

$$RPC = 1.96 \times \sqrt{\frac{\sum(y_2 - y_1)^2}{n}} \quad [9]$$

Since an LLD >5 mm could be associated with an increased risk of osteoarthritis of the knee or hip, the sensitivity and specificity defined below were also calculated to evaluate automatic LLD measurement:

$$sensitivity = \frac{True \ Positive}{True \ Positive + False \ Negative} \quad [10]$$

$$specificity = \frac{True \ Negative}{True \ Positive + False \ Negative} \quad [11]$$

Results

Comparison of segmentation with alternative DL models

The comparison of segmentation between the proposed cascaded LLDNet and the alternative DL models of ResUNet (12), TransUNet (20), and the single LLDNet is summarized in *Table 2*. Experimental results demonstrated that the cascaded LLDNet achieved the best segmentation accuracy in terms of similarity and stability indices. The visualized comparison of the segmentation results is also displayed in *Figure 5*, which further demonstrates that the cascaded LLDNet could greatly decrease segmentation errors.

Furthermore, although the similarity indices achieved by the cascaded LLDNet were only slightly improved in quantity (0.5–1% percentage increase of indices; P<0.05; paired Wilcoxon signed-rank test), the stability indices improved dramatically between the different DL models under comparison (13–47% percentage decrease of indices: P<0.05, paired Wilcoxon signed-rank test), as shown in *Table 2*. The percentage improvement in terms of similarity and stability indices shown in *Table 2* not only demonstrate the superiority of the cascaded LLDNet in segmenting pediatric legs, but also emphasize the necessity of the comprehensive assessment in the LLD study.

Comparison with radiology reports

Other than the evaluation of segmentation accuracy, the consistency in the measurement of leg length between DL

Table 2 Comprehensive evaluation of pediatric leg segmentation by 4 similarity indices and 5 stability indices (mean ± SD)

Deep learning model	Similarity indices				Stability indices				
	Dice	Jaccard	Precision	Recall	MD	HD	HD95	ASSD	RMSD
ResUNet	0.9754±0.01*	0.9523±0.02*	0.9735±0.02*	0.9776±0.01	0.2693±0.22*	5.7778±8.90*	1.2044±1.52*	0.2561±0.14*	0.6634±0.62*
TransUNet	0.9720±0.01*	0.9458±0.02*	0.9679±0.02*	0.9763±0.01	0.2708±0.14*	3.8934±4.03*	1.1963±1.36*	0.2790±0.12*	0.6120±0.40*
Single LLDNet	0.9759±0.01*	0.9530±0.02*	0.9747±0.01*	0.9772±0.01	0.2554±0.25*	4.3214±6.70*	1.2048±1.90*	0.2483±0.15*	0.6097±0.62*
Cascaded LLDNet	0.9772±0.01	0.9555±0.01	0.9783±0.01	0.9762±0.01	0.2195±0.07	3.0277±2.81*	1.0455±0.28*	0.2255±0.07	0.5119±0.19
The maximum percentage improvement	↑0.5%	↑1.0%	↑1.0%		↓18%	↓47%	↓13%	↓19%	↓22%

The ResUNet, TransUNet, single LLDNet, and proposed cascaded LLDNet were compared. *, the cascaded LLDNet significantly improved the performance according to paired Wilcoxon signed-rank test. SD, standard deviation; MD, mean distance; HD, Hausdorff distance; ASSD, average symmetric surface distance; RMSD, root mean square deviation; LLD, leg length discrepancy.
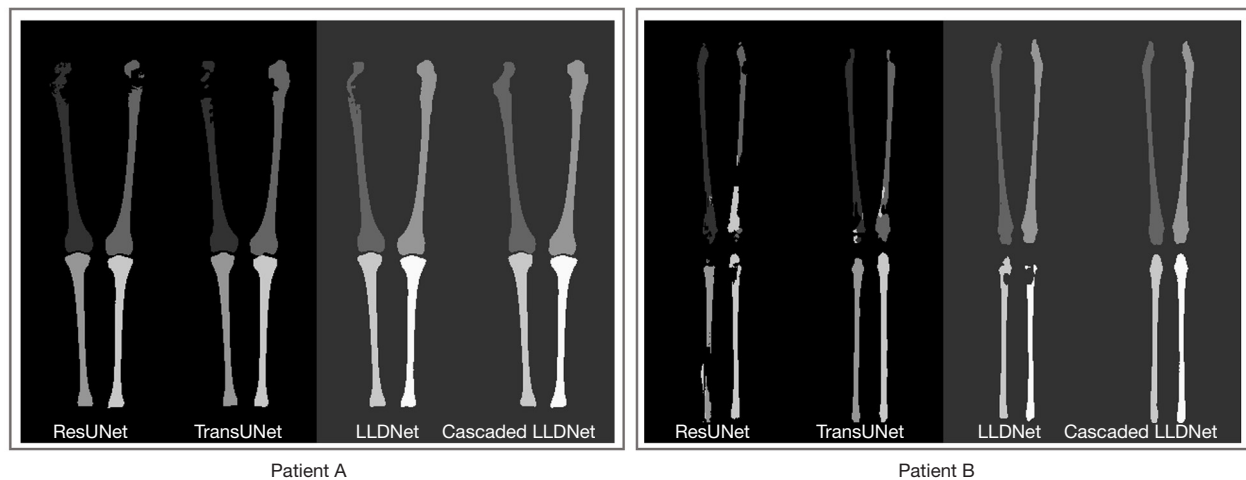
Figure 5 Comparison of segmentation between different deep learning models on patients A and B, including the ResUNet, TransUNet, single LLDNet, and cascaded LLDNet.

and ground truth was also comprehensively assessed by the 5 metrics of PCC, MSE, MAE, MaxAE, and CCC. The 5 metrics were also compared between the proposed cascaded LLDNet with the alternative DL models of the ResUNet (12), TransUNet (20), and the single LLDNet (*Table 3*).

Regarding the comparison of the measurement of the length of the femur, tibia, and full leg, as shown in *Table 3*, the proposed cascaded LLDNet achieved better performance in the measurement of consistency over alterative DL models under comparison. Regarding LLD measurement, the cascaded LLDNet also obtained the best consistency with ground truth (PCC =0.94, P<0.05; MSE =0.02, MAE =0.09, MaxAE =1.85, CCC =0.94) over the other methods.

In addition, *Figure 6* exhibits the incidence of different anatomic LLD magnitudes among the participants, where LLDs of 0–2 and 8–10 mm had the highest and lowest incidence in the population in our study, respectively. With respect to the incidence of different anatomic LLD magnitudes among the participants, the scatter plots (with PCC, CCC, MSE, MAE, and MaxAE) of the comparison between the radiology report and DL models (*Figure 7*), indicated that the best measurement consistency of femoral length, tibial length, leg length, and LLD was achieved by the cascaded LLDNet. The Bland-Altman plot (*Figure 8*) was also used to assess the agreement between the manual measurement and the leg length calculated by the cascaded LLDNet. Regarding the LLD difference between the

Table 3 Comparison between the radiology reports and the automated measurements in terms of Pearson correlation coefficient, mean squared error, mean absolute error, maximum absolute error, and concordance correlation coefficient

Anatomical structure	Deep learning model	Pearson correlation coefficient	Mean squared error (cm)	Mean absolute error (cm)	Maximum absolute error (cm)	Concordance correlation coefficient
Femur	ResUNet	0.93*	6.09	0.57	35.9	0.92
	TransUNet	0.91*	7.46	0.68	27.4	0.90
	Single LLDNet	0.95*	3.99	0.42	28.3	0.94
	Cascaded LLDNet	0.99*	0.07	0.19	1.21	0.99
Tibia	ResUNet	0.89*	4.89	0.49	34.4	0.89
	TransUNet	0.89*	5.89	0.59	22.5	0.88
	Single LLDNet	0.91*	3.84	0.41	28.4	0.91
	Cascaded LLDNet	0.99*	0.06	0.18	1.59	0.99
Full leg length	ResUNet	0.99*	0.50	0.22	8.91	0.99
	TransUNet	0.99*	0.63	0.24	14.3	0.99
	Single LLDNet	0.99*	0.16	0.14	6.13	0.99
	Cascaded LLDNet	0.99*	0.02	0.10	1.71	0.99
LLD	ResUNet	0.59*	0.31	0.19	6.14	0.53
	TransUNet	0.50*	0.49	0.23	7.08	0.41
	Single LLDNet	0.70*	0.18	0.15	4.62	0.67
	Cascaded LLDNet	0.94*	0.02	0.09	1.85	0.94

The ResUNet, TransUNet, single LLDNet, and proposed cascaded LLDNet were compared. *, significant correlation with P<0.05 between radiology reports and the automated measurements. LLD, leg length discrepancy.
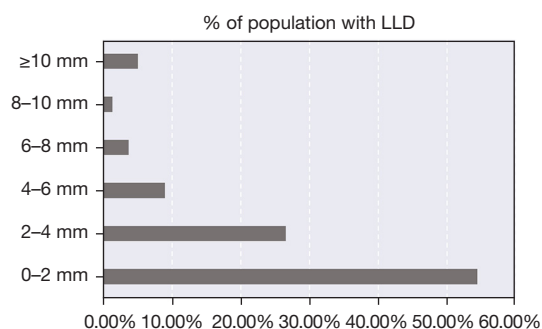
Figure 6 Incidence of anatomic LLD magnitude. LLD, leg length discrepancy.

manual and automatic measurements, the Bland-Altman analysis showed a mean bias ± SD of 0.003±0.15 cm and RPC of 0.29.

Using the results of the manual measurement of LLD as the gold standard, we computed the sensitivity and specificity of the automatic LLD measurements of all cases with LLD >5 mm and LLD >10 mm. The sensitivity and specificity established for LLD >5 mm were 0.792 and 0.962, respectively, while those for LLD >10 mm were 0.938 and 0.992, respectively.

Discussion

In the present study, we developed a cascaded LLDNet for automatic LLD measurement and then comprehensively assessed its performance on a radiographic dataset covering children at all stages from infancy to adolescence and with a wide range of diagnoses. The LLDNet performed better in the LLD study in terms of similarity and stability compared to the other models (*Table 2* and *Figure 5*). Experiments comprehensively employing a variety of metrics and statistical analyses demonstrated the superiority of the cascaded LLDNet in the LLD study over ResUNet (12), TransUNet (20), and the single LLDNet (*Figure 3A*).

Previous research on this subject has only been carried out on children from a preschool stage (13) and in a
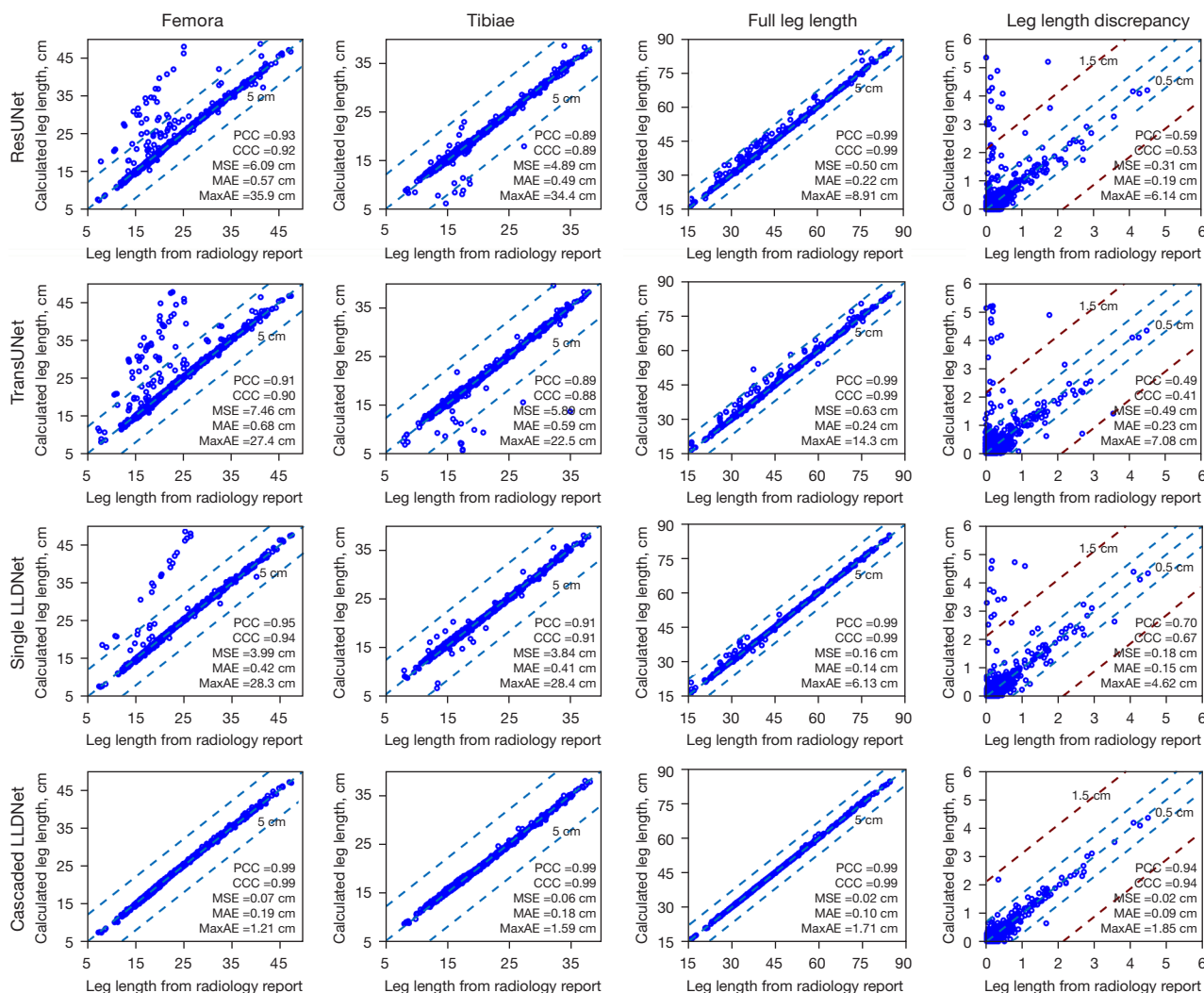
Figure 7 Scatter plots of the radiology report and deep learning models. The first to fourth rows represent results achieved by the ResUNet, TransUNet, single LLDNet, and cascaded LLDNet, respectively. The first to fourth rows columns represent the femora, tibiae, full leg length, and LLD, respectively. PCC, Pearson correlation coefficient; CCC, concordance correlation coefficient; MSE, mean squared error; MAE, mean absolute error; MaxAE, maximum absolute error; LLD, leg length discrepancy.

relatively small cohort (12); however, children with LLD at infant and toddler stages represent a particular challenge in clinical routines. Children experience many remarkable changes in terms of both the structure and alignment of their legs during development (14), as displayed in *Figure 1*. Therefore, a unified solution using DL on a large and comprehensive radiographic dataset covering children at all stages, from infancy to adolescence, and with a wide range of diagnoses, is desirable. Therefore, the present study employed a large sample number of 972 cases of pediatric bilateral lower limb radiographs to develop a unified solution for the LLD study and to avoid statistical

deviations. Although a variety of different types of samples were enrolled, the metrics (*Table 3*) achieved by the cascaded LLDNet demonstrated a better consistency compared to those reported in previous radiology studies (12) and (13) when measuring the full leg length and LLD.

A comprehensive assessment is essential in an LLD study. The previous study (12) which used the same-leg segmentation-based strategy, only adopted the spatial overlap of the Dice similarity coefficient to evaluate segmentation accuracy. However, the Dice values only changed slightly between different DL models, as displayed in *Table 2*, indicating an insufficient ability to assess the
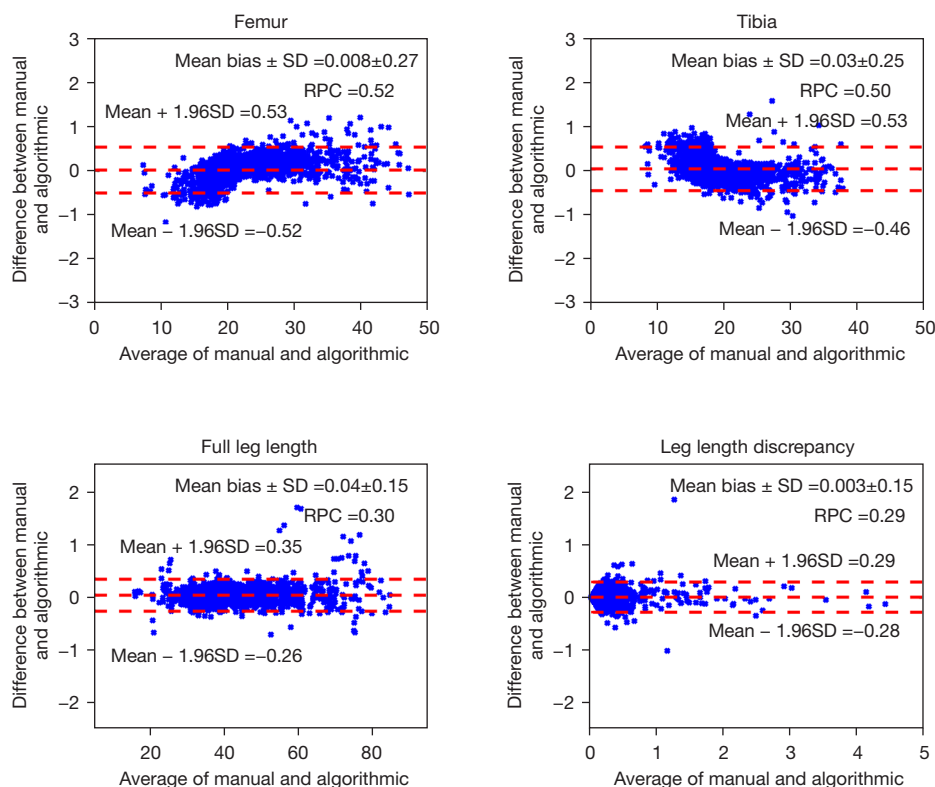
Figure 8 Bland-Altman plot to assess the agreement between the reference leg lengths and the leg length that was calculated through the cascaded LLDNet. RPC, coefficient of repeatability; SD, standard deviation; LLD, leg length discrepancy.

segmentation. *Figure 5* further demonstrates the limitation of the Dice index in measuring different segmentations. Our present study employed a comprehensive assessment strategy, including Dice, Jaccard, Precision, Recall, MD, HD, HD95, ASSD, and RMSD in terms of both segmentation similarity and stability, and we could comprehensively evaluate the improvement of different segmentation models. For clarity, the maximum percentage improvement of the cascaded LLDNet over other DL methods is summarized in the last line of *Table 2*. Although there was only a slight improvement in the similarity indices (0.5–1% in percentage increase), a dramatic improvement in the stability indices was observed (13–47% percentage decrease), which demonstrates the necessity of a comprehensive assessment in LLD study.

Knutson *et al.* (22) reported that only 10% of the average population has exactly equal leg length, 90% of patients have at least a 1 mm LLD, approximately 50% of the population has a 4 mm LLD, and approximately 90% of the population has an LLD of 10 mm or less. There is evidence that an LLD of >5 mm may be associated with

a high risk of hip, knee, or back problems (2,3). These patients tend to adapt to the LLD over a long period and temporarily experience a reprieve of symptoms because of the body's compensatory ability, but the LLD gradually increases with age, which is especially obvious in children (23,24). Thus, in our study, we compared the sensitivity and specificity of automatic LLD measurements of LLD >5 mm in all of these cases, resulting in a sensitivity of 0.792 and a specificity of 0.962. For comparison, we also calculated the sensitivity and specificity of the automatic measurement of LLD >10 mm, achieving a higher sensitivity of 0.938 and specificity of 0.992 as compared to the sensitivity of 0.88 and specificity of 0.97 reported in a study (13).

Some limitations to our study should be noted. First, this research was not grouped by children's age and did not evaluate the degree of ossification of each ossification center according to the different growth stages, which might have introduced errors in automatic segmentation. In our follow-up research, we will conduct a study of details more relevant to the age group of the children and further evaluate whether the different degrees of ossification in

the ossification center affect the automatic segmentation and measurement results. Second, the data included in the present study were only collected from 2 locally limited devices, the DigitalDiagnost 3 and the SONIALVISION Plus, and did not include other machines and sites, such as the ultra-low dose X-ray imaging system (EOS), which performs full-length, lower-extremity radiographs of the upright position (13,25) and may generate more accurate results. However, the EOS imaging system was not equipped for use due to the limitation of the practical conditions of the study. As a result, we used 2 existing photography systems for acquiring full-length, lower-extremity radiographs of the supine position (patients with internal/external fixation devices were excluded). Thus, the effectiveness and robustness of our method should be further validated in future studies with external, multisite data for higher clinical applicability. Third, although the algorithm we used to measure the femur and tibia was accurate, slight errors are unavoidable. The calculation of the full leg length was based on the sum of the length of the femur and tibia, while the calculation of LLD was based on the result of the full leg length, which could have generated a quadratic error and results with a relatively low prediction accuracy of LLD. Finally, an independent remeasurement of leg length was not performed.

We conducted a comprehensive study of LLD in a large sample comprising a diversity of case types. The LLD study consisted of 4 stages of implementation: cascaded LLDNet for leg segmentation, postprocessing of segmentation masks, anatomical landmark identification, and leg length measurement. Experimental results demonstrated that, compared with other LLD studies, our method was able to achieve a superior performance based on a comprehensive evaluation in terms of similarity, stability, and consistency with radiology reports.

Acknowledgments

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-22-282/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-22-282/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by institutional ethics board of Xiamen Children's Hospital, and individual consent for this retrospective analysis was waived. There are no details on persons mentioned within the text.

References

1. Khamis S, Carmeli E. A new concept for measuring leg length discrepancy. J Orthop 2017;14:276-80.
2. Murray KJ, Molyneux T, Le Grande MR, Castro Mendez A, Fuss FK, Azari MF. Association of Mild Leg Length Discrepancy and Degenerative Changes in the Hip Joint and Lumbar Spine. J Manipulative Physiol Ther 2017;40:320-9.
3. Tallroth K, Ristolainen L, Manninen M. Is a long leg a risk for hip or knee osteoarthritis? Acta Orthop 2017;88:512-5.
4. Desai AS, Dramis A, Board TN. Leg length discrepancy after total hip arthroplasty: a review of literature. Curr Rev Musculoskelet Med 2013;6:336-41.
5. Vogt B, Gosheger G, Wirth T, Horn J, Rödl R. Leg Length Discrepancy- Treatment Indications and Strategies. Dtsch Arztebl Int 2020;117:405-11.
6. Sabir AB, Faizan M, Ishtiaq M, Jilani LZ, Ahmed S, Shaan ZH. Limb length discrepancy after total knee arthroplasty: Unilateral versus bilateral, a comparative study at tertiary centre. J Clin Orthop Trauma 2020;11:S740-5.
7. Gallo MC, Chung BC, Tucker DW, Piple AS, Christ AB, Lieberman JR, Heckmann ND. Limb Length Discrepancy

in Total Hip Arthroplasty: Is the Lesser Trochanter a Reliable Measure of Leg Length? J Arthroplasty 2021;36:3593-600.

8. Chua CXK, Tan SHS, Lim AKS, Hui JH. Accuracy of biplanar linear radiography versus conventional radiographs when used for lower limb and implant measurements. Arch Orthop Trauma Surg 2022;142:735-45.

9. White SA, Shellikeri S, Muñoz ML, Edgar JC, Nguyen JC, Sze RW. Can Radiology Technologists be Trained to Measure Leg Length Discrepancies as Accurately as Pediatric Radiologists? Acad Radiol 2022;29:51-5.

10. Gao Y, Liu B, Zhu Y, Chen L, Tan M, Xiao X, Yu G, Guo Y. Detection and recognition of ultrasound breast nodules based on semi-supervised deep learning: a powerful alternative strategy. Quant Imaging Med Surg 2021;11:2265-78.

11. Mallio CA, Quattrocchi CC, Beomonte Zobel B, Parizel PM. Artificial intelligence, chest radiographs, and radiology trainees: a powerful combination to enhance the future of radiologists? Quant Imaging Med Surg 2021;11:2204-7.

12. Zheng Q, Shellikeri S, Huang H, Hwang M, Sze RW. Deep Learning Measurement of Leg Length Discrepancy in Children Based on Radiographs. Radiology 2020;296:152-8.

13. Tsai A. Anatomical landmark localization via convolutional neural networks for limb-length discrepancy measurements. Pediatr Radiol 2021;51:1431-47.

14. Alshryda S, Jackson L, Thalange N, AlHammadi A. Pediatric Orthopedics for Primary Healthcare. 1 ed. Springer International Publishing; 2021.

15. Stewart RJ, Patterson CC, Mollan RA. Ossification of the normal femoral capital epiphysis. J Bone Joint Surg Br 1986;68:653.

16. Sugawara R, Watanabe H, Taki N, Aihara T, Furukawa R, Nakata W, Takeshita K, Kikkawa I. New radiographic standards for age at appearance of the ossification center of the femoral head in Japanese: Appearance at ≤12 months of age is normal in Japanese infants. J Orthop Sci 2019;24:166-9.

17. Garn SM, Rohmann CG, Silverman FN. Radiographic standards for postnatal ossification and tooth calcification. Med Radiogr Photogr 1967;43:45-66.

18. Hedequist D, Heyworth BE. Pediatric femur fractures: A parctical guide to evaluation and management. Springer, Boston, MA; 2016.

19. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. European Conference on Computer Vision (ECCV), Munich, Germany 2018:11211.

20. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:210204306 2021.

21. Zheng Q, Wu Y, Fan Y. Integrating Semi-supervised and Supervised Learning Methods for Label Fusion in Multi-Atlas Based Image Segmentation. Front Neuroinform 2018;12:69.

22. Knutson GA. Anatomic and functional leg-length inequality: a review and recommendation for clinical decision-making. Part I, anatomic leg-length inequality: prevalence, magnitude, effects and clinical significance. Chiropr Osteopat 2005;13:11.

23. Gordon JE, Davis LE. Leg Length Discrepancy: The Natural History (And What Do We Really Know). J Pediatr Orthop 2019;39:S10-3.

24. Wynes J, Schupp A. Assessment of Pediatric Limb Length Inequality. Clin Podiatr Med Surg 2022;39:113-27.

25. Tsai A. A deep learning approach to automatically quantify lower extremity alignment in children. Skeletal Radiol 2022;51:381-90.