International Journal of Computer Assisted Radiology and Surgery (2024) 19:939–950
https://doi.org/10.1007/s11548-024-03089-z

ORIGINAL ARTICLE

PELE scores: pelvic X-ray landmark detection with pelvis extraction and enhancement

Zhen Huang[1,3] · Han Li[2,3] · Shitong Shao[4] · Heqin Zhu[2,3] · Huijie Hu[1,3] · Zhiwei Cheng[5] · Jianji Wang[6] · S. Kevin Zhou[2,3,7]

Received: 17 September 2023 / Accepted: 27 February 2024 / Published online: 15 March 2024
© CARS 2024

Abstract

Purpose Pelvic X-ray (PXR) is widely utilized in clinical decision-making associated with the pelvis, the lower part of the trunk that supports and balances the trunk. In particular, PXR-based landmark detection facilitates downstream analysis and computer-assisted diagnosis and treatment of pelvic diseases. Although PXR has the advantages of low radiation and reduced cost compared to computed tomography (CT), it characterizes the 2D pelvis-tissue superposition of 3D structures, which may affect the accuracy of landmark detection in some cases. However, the superposition nature of PXR is implicitly handled by existing deep learning-based landmark detection methods, which mainly design the deep network structures for better detection performances. Explicit handling of the superposition nature of PXR is rarely done.
Methods In this paper, we explicitly focus on the superposition of X-ray images. Specifically, we propose a pelvis extraction (PELE) module that consists of a decomposition network, a domain adaptation network, and an enhancement module, which utilizes 3D prior anatomical knowledge in CT to guide and well isolate the pelvis from PXR, thereby eliminating the influence of soft tissue for landmark detection. The extracted pelvis image, after enhancement, is then used for landmark detection.
Results We conduct an extensive evaluation based on two public and one private dataset, totaling 850 PXRs. The experimental results show that the proposed PELE module significantly improves the accuracy of PXRs landmark detection and achieves state-of-the-art performances in several benchmark metrics.
Conclusion The design of PELE module can improve the accuracy of different pelvic landmark detection baselines, which we believe is obviously conducive to the positioning and inspection of clinical landmarks and critical structures, thus better serving downstream tasks. Our project has been open-sourced at https://github.com/ECNUACRush/PELEscores.

Keywords Bone extraction · CT structural knowledge · Landmark detection · Pelvis X-rays

Introduction

Pelvis, the foundation of the human torso, is of great significance to human health [24]. In clinical decision-making, 2D pelvic X-rays (PXRs) are widely used due to their low radiation exposure and low cost [30, 34]. To take advantage of PXRs and improve their clinical utility, research is dedicated to automatic analysis of PXRs [6, 8, 34]. For instance, PXR landmark detection, aiming to localize some key anatomi-

Zhen Huang and Han Li have contributed equally to this work.

This work was supported by Natural Science Foundation of China under Grant 62271465 and Open Fund Project of Guangdong Academy of Medical Sciences, China (No. YKY-KF202206).

✉ S. Kevin Zhou
skevinzhou@ustc.edu.cn

1 Computer Science Department, University of Science and Technology of China (USTC), Hefei 230026, China

2 School of Biomedical Engineering, Division of Life Sciences and Medicine, USTC, Hefei 230026, China

3 Center for Medical Imaging, Robotics, Analytic Computing and Learning (MIRACLE), Suzhou Institute for Advanced Research, USTC, Suzhou 215123, China

4 Southeast University, Nanjing 210018, China

5 Z2Sky Technologies Inc., Suzhou 215123, China

6 Affiliated Hospital of Guizhou Medical University, Guiyang 550000, China

7 Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

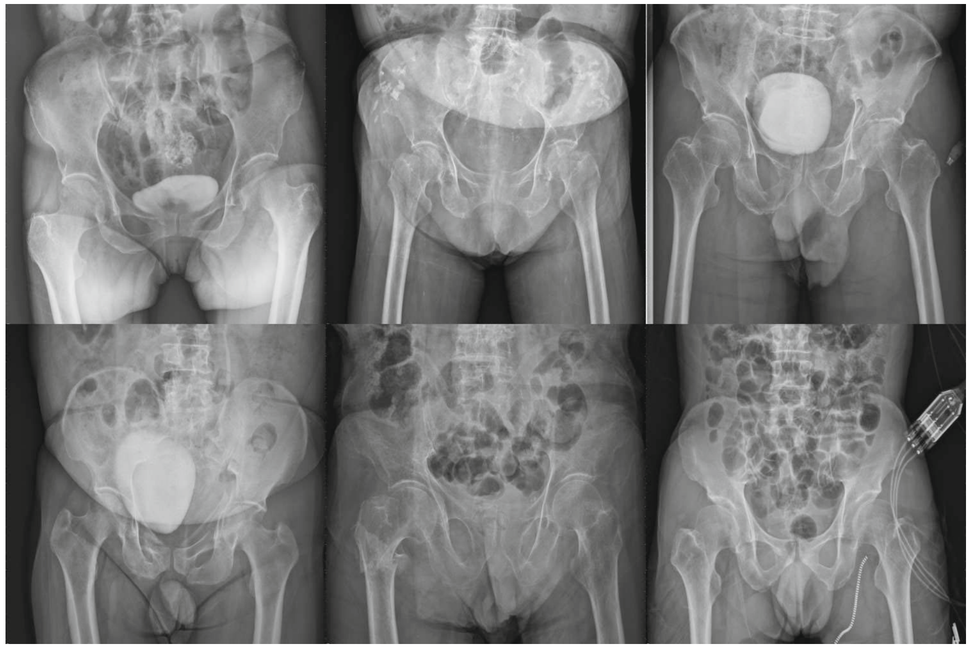

Fig. 1 Pelvic X-ray poses the superposition problem, which renders a severe difficulty in finding landmarks

cal points, plays an essential role in clinical computer-aided diagnosis and treatment [25, 36, 41]. On the one hand, landmark detection can automatically offer key points for clinical use [4, 20–22]; on the other hand, it can be used for some downstream analysis tasks, including segmentation [5, 31, 38], registration [12, 26], pathological diagnosis [9, 37], and surgical planning [33].

Mainstream X-ray landmark detection methods show powerful performances based on elegant deep network designs [1, 7, 18, 27, 40]. These methods mainly focus on (1) designing, for better learning of discriminative features, novel network architectures, such as spatial configuration network [28], patch-based iterative network [19], graph-based convolutional neural network based on Chebyshev polynomials [10], and context encoding-constrained network [23]; and (2) developing better landmark representation and decision-making strategies, such as multi-channel heatmaps [39] and local voting [35]. However, these methods are still challenged by the so-called pelvis-tissue superposition, that is, the pelvis can be covered by soft tissues like large and small intestines, bladder, and urethra (as in Fig. 1), as they implicitly handle the superposition, which is exemplified by training images. However, the image variations posed by superposition are so diverse that it is difficult to exhaust them with limited training data.

In this paper, we take an explicit approach and attempt to extract the pelvis from a PXR before landmark detection, therefore addressing the pelvis-tissue superposition challenge upfront. This is shown in Fig. 2. Specifically, we first design a PELvis Extraction (PELE) module based on structural priors in 'unpaired' 3D CT images to extract the pelvis bone from PXRs. The extracted pelvis is further fed into an enhancement module for extra quality improvement. The final enhanced pelvis serves as the input of definitive landmark detection for improved detection accuracy. Our contributions are as follows:

- To the best of our knowledge, we are the first to focus on the pelvis-tissue superposition challenge explicitly to benefit pelvic landmark detection. We propose to relieve this challenge by performing explicit pelvis extraction before feeding the image into a landmark detection network.
- The significance of PELE has been demonstrated by quantitative and qualitative results at different data scales with different baselines. State-of-the-art (SOTA) pelvic landmark detection performances have been obtained on two public datasets and an in-house real-patient dataset. The landmark detection errors are significantly reduced with the aid of PELE.

Method

Pelvis extraction (PELE)

As shown in Fig. 3, the PELE module consists of two stages: (a) the image decomposition stage via F_{DE} and (b) the domain adaptation via F_{DA}.

Image Decomposition. Ideally, we aim to decompose a 2D PXR X_{raw} into a pelvis-only image X_{bone} and a tissue-only image X_{tissue}, i.e., $X_{\mathrm{raw}} \rightarrow (X_{\mathrm{bone}}, X_{\mathrm{tissue}})$ by a learned

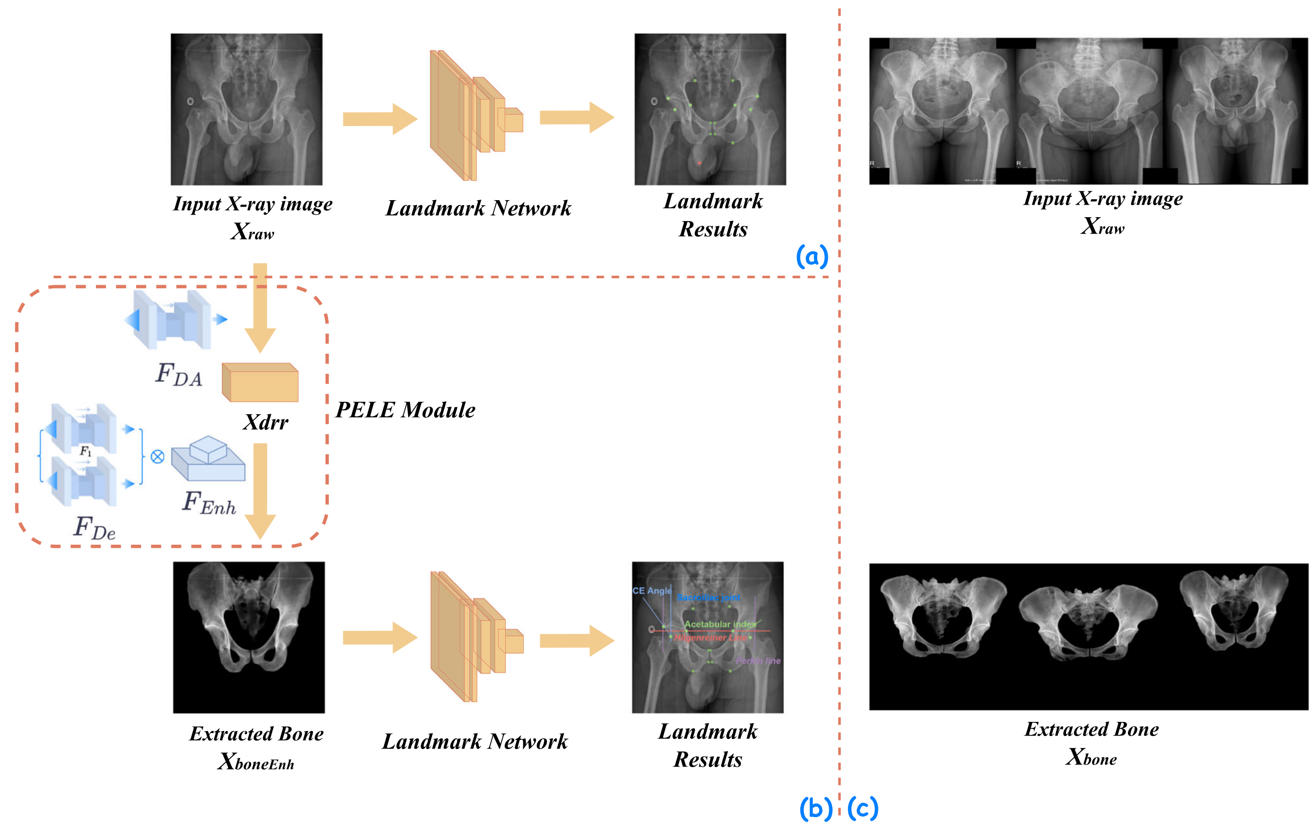

Fig. 2 The existing landmark detection method may have detection results with significant errors because of the pelvis-tissue supposition challenge. We break this challenge using the proposed pelvis extraction (PELE) module, and then the extracted pelvis image is used for accurate landmark detection. **a** Process of normal landmark detection. **b** Process of landmark detection using PELE module. Note that F_{De}, F_{DA}, F_{Enh} and Landmark Network, respectively, refer to (a), (b), (c), and (d) in Fig. 3. **c** Some comparison renderings before and after using the PELE module

deep network. However, such learning is challenging since no pelvis-only images are paired with 2D PXRs in common datasets. Therefore, inspired by [2, 13, 16, 17], we resort to 3D prior knowledge in 3D CT and use a 2D digitally reconstructed radiograph (DRR) image to bridge the gap between 3D CT and 2D PXR.

From a 3D CT volume V_{raw} with its isolated pelvis part V_{bone}, we first create 2D DRR images of V_{raw} and V_{bone}, denoted as I_{raw} and I_{bone}, respectively, using the DeepDRR [32] algorithm. Then we learn a deep neural network F_{DE} to perform DRR-based decomposition:

$$I_{\mathrm{bone}} = F_{\mathrm{DE}}(I_{\mathrm{raw}}), \tag{1}$$

based on an L_1 loss function.

To boost the performance of F_{DE}, we introduce bone mask segmentation as an auxiliary task in addition to bone extraction. Specifically, in the training dataset, a CT volume V_{raw} is associated with the bone mask V_{mask} annotation; thus the bone part is obtained via $V_{\mathrm{bone}} = V_{\mathrm{raw}} \odot V_{\mathrm{mask}}$, where $\odot$ refers to the element-wise product. We project 3D V_{mask} to create the projected 2D mask image I_{mask} using DeepDRR. These images are used in a two-stage operation: (1) First, we learn two networks, an nnU-Net F_1 and a U-Net F_2. Both take the 2D DRR image I_{raw} as input, the nnU-Net outputs I_{mask}, and the U-Net outputs I_{bone}; (2) Then, we take their product $I_{\mathrm{mask}} \odot I_{\mathrm{bone}}$ as the final bone prediction $\hat{I}_{\mathrm{bone}}$.

$$\begin{aligned} I_{\mathrm{mask}} &= F_1(I_{\mathrm{raw}});\ I_{\mathrm{bone}} = F_2(I_{\mathrm{raw}});\ \hat{I}_{\mathrm{bone}} \\ &= F_{\mathrm{DE}}(I_{\mathrm{raw}}) = I_{\mathrm{mask}} \odot I_{\mathrm{bone}}. \end{aligned} \tag{2}$$

Domain Adaptation. The DRR image has the same size as the PXR, but there is a domain gap between DRR and PXR; hence, the decomposition model F_{DE} cannot be directly applied to the PXR X_{raw}. To further narrow the domain gap between them, we resort to domain adaptation (DA). Specifically, we utilize the CycleGAN [42] as our DA backbone, which learns a forward mapping network F_{DA} from the PXR to the DRR and an auxiliary inverse mapping $I_{\mathrm{raw}} = F_{\mathrm{DA}}(X_{\mathrm{raw}})$. Finally, the predicted pelvis for X_{raw},

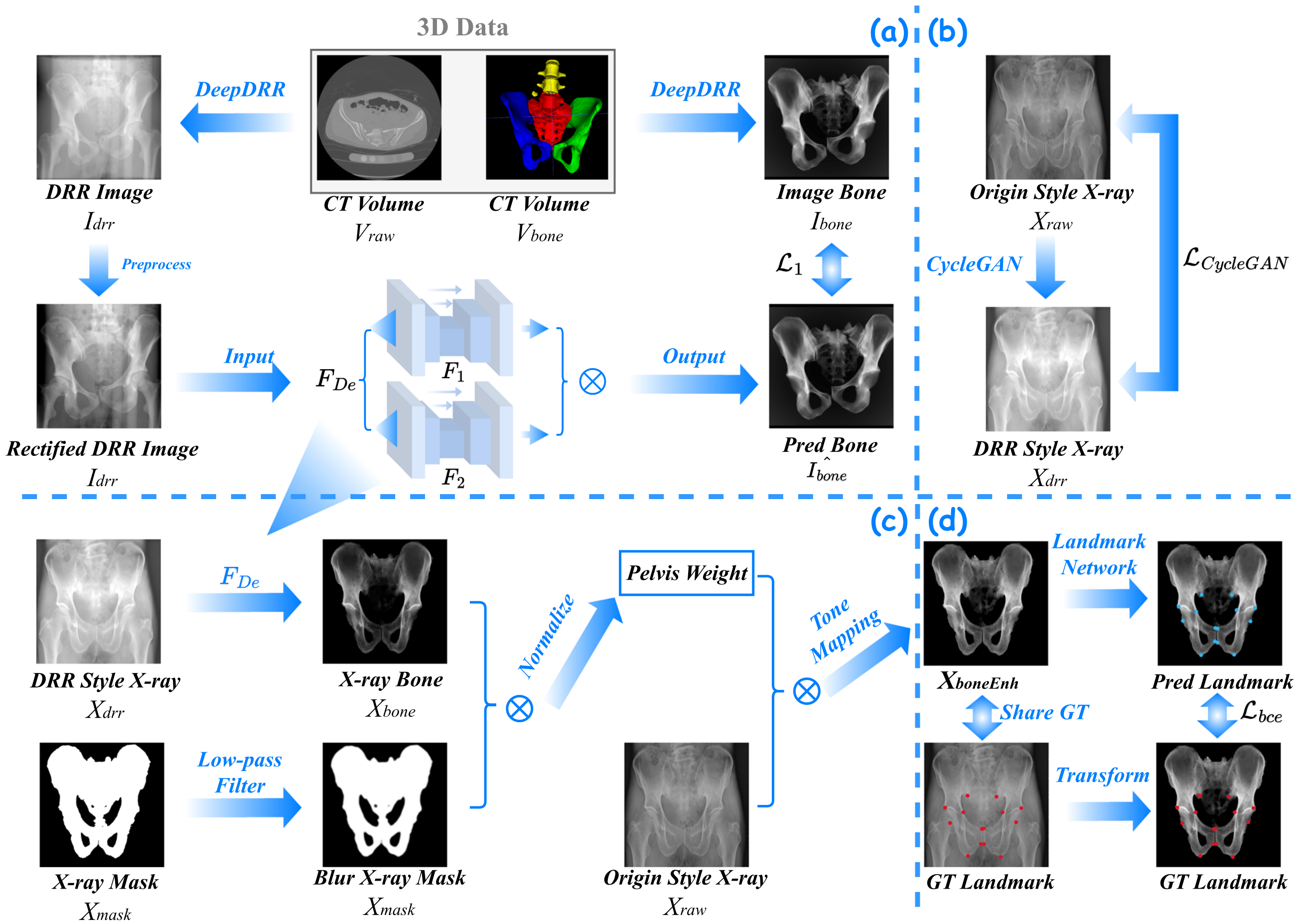

Fig. 3 **a**, **b** The diagram of the proposed PELvis Extraction (PELE) module. **c** The enhancement module. **d** The landmark detection flow

denoted by X_{bone}, is given by

$$X_{\text{bone}} = F_{\text{DE}}(\hat{I}_{\text{bone}}) = F_{\text{DE}}(F_{\text{DA}}(X_{\text{raw}})). \quad (3)$$

Pelvis enhancement

The X_{bone} generated from F_{DE} sometimes contains artifacts, especially in the ‘dark’, less-penetrated areas like hip bones, sacrum bones, and tail bones, which affect the subsequent diagnosis. To solve this, we propose an enhancement module to obtain the final result, X_{boneEnh}, as in Fig. 3c. Firstly, we smooth the transition and normalization of the pelvic contour edge in the X_{bone} to get processed pelvic bone X_{bonePre}; we use the Gaussian filter as a low-pass filter for a smooth transition.

$$X_{\text{bonePre}} = \mathcal{N}\left[\mathcal{G}\left(X_{\text{mask}}\right) \odot X_{\text{bone}}\right], \quad (4)$$

where X_{bone} denotes the pelvis extracted from X_{raw} before, X_{mask} denotes the binary mask of bone, $\mathcal{N}$ denotes the normalization operator, and $\mathcal{G}$ denotes a low-pass filter operation (e.g., Gaussian filter). Then we multiply the X_{bonePre} with X_{raw} to obtain the image containing PXR details and perform the tone mapping operations (e.g., Gamma correction) to compress the dynamic range of dark areas to show the details clearly. The final enhanced image is X_{boneEnh}.

Pelvic landmark detection

As shown in Fig. 3d, we use the enhanced pelvis image X_{boneEnh} to train a landmark detection network Φ. Firstly, we annotate landmark labels on X_{raw}, which are used to generate the ground truth (GT) heatmaps H_{gt}. Given an annotated landmark position (x_0, y_0), its GT heatmap $H_{\text{gt}}(x, y)$ is formulated as

$$H_{\text{gt}}(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma^2}\right\}, \quad (5)$$

where σ represents the standard deviation of the Gaussian distribution, which dictates the shape and intensity values of the heatmap. As a critical hyperparameter, we empirically determine its optimal value to be 3 based on our experiments and analysis. Then, we train the detection network with X_{boneEnh} supervised by GT heatmaps H_{gt}. As X_{boneEnh} and X_{raw} share the same imaging grid, the predicted landmarks can be directly applied to the original image X_{raw}.

This work, uses two landmark baselines, U-Net [29], and GU2Net [41]. U-Net [29] is a popular baseline in medical analysis, and GU2Net [41] is a universal model for landmark detection, which combines local features and global context information. We use the cross-entropy loss to optimize the above networks. After obtaining the predicted heatmap, we apply arg max to extract coordinates of the strongest response as predicted landmarks.

Experiment

Settings

Datasets. For CT datasets, we use the public dataset CTPelvic1K [24], which contains 1184 volumes (over 320K slices). We use the DeepDRR [32] algorithm to generate DRR images and resize them to 512 × 512 based on a bilinear interpolation. Finally, we generate 919 DRR images to train F_{DA}, and select 200 high-quality images to train F_{DE}. For PXRs, we curate 850 images from three sources. Dataset1: open-source CGMH-PelvisSeg, provided by CGMHai Lab,[1] which contains 400 high-resolution PXRs. Dataset2: A 150-image open-source dataset provided by [8].[2] Dataset3: An in-house dataset of 300 images we retrospectively collect from cooperative hospitals with the institutional review board (IRB) policies of contributing sites. All data are under Creative Commons license CC-BY-NC-SA at least, and we will keep the license unchanged. To handle varying sizes of these images, we resize them to 512 × 512.

Implementation Details. The decomposition network F_{DE} consists of nnU-Net and U-Net. For domain adaptation network F_{DA}, the generators G_{D} and G_{X} both use a 9-block ResNet, while D_{D} and D_{X} use PatchGAN [42]. For the decomposition network F_{DE}, we train 1000 epochs using the basic setup of nnU-Net [14]. For U-Net [29], we use Adam optimizer with an initial learning rate of 2e−4 and a batch size of 16. F_{DA} is trained for 200 epochs; reasonable results are obtained after 100 epochs. λ_{cycle} is increased to 15 to reduce the data limitation's effect and obtain good results. Both F_{DE} and F_{DA} are implemented based on an RTX3090 GPU. Training F_{DE} takes about 20 h, and training F_{DA} takes approximately 30 h. For landmark detection, we refer to the setting of [41]: 3×3 convolution followed by batch normalization and leaky RELU activation. The batch size is set to 4, the learning rate (LR) is [1e−4, 1e−2], and a cyclic scheduler is used to decrease LR within this interval. Binary cross-entropy (BCE) loss and Adam optimizer are used. The dataset is randomly divided into 7/1/2 for train/val/test.

The extraction performance of the PELE module

We first test the decomposition performance of the PELE module. We use commonly used metrics for segmentation evaluation: Structural Similarity (SSIM) to measure the structural consistency, Learned Perceptual Image Patch Similarity (LPIPS) to measure image block similarity, Dice Score and Hausdorff Distance (the 95th percentile or HD95) to measure the effectiveness of the extracted bone from the PELE module. The pelvic areas of 100 PXRs are manually marked using LabelMe[3] by a medical imaging practitioner with five years of experience. The marking approach involves an approximation technique where polylines are used to represent curves.

We compare two baselines: U-Net and CycleGAN. The U-Net takes a PXR X_{raw} as an input directly into the decomposition F_{DE} network to obtain the extracted bone X_{bone}. The CycleGAN, different from Eq. (3), directly trains a network that maps between X_{raw} and X_{bone}. Note that X_{bone} are not readily available. Our previous experiments indicate that among the various metrics evaluated, PSNR (peak signal-to-noise ratio) shows the strongest correlation with image quality. Consequently, we utilize PSNR as the primary criterion for sorting images generated by PELE. Subsequently, a medical imaging practitioner with five years of experience selects 297 PXRs, prioritizing image quality and particularly focusing on the completeness of the images, to align with the number of training sets used in our experiment.

From Table 1, we perceive that PELE outperforms the baseline by a large margin: SSIM and LPIPS reach 0.923 and 0.083, respectively. In Fig. 4 as well as in Fig. 2c, we show the visualization of different modules: The bone is well separated from the soft tissue in PELE. At the same time, comparing results before and after using the enhancement module, F_{Enh}, we observe that less-penetrated areas such as hip bones, sacrum bones, and tail bones are more visible, which proves that F_{Enh} is also instrumental.

In order to further evaluate whether the bone extraction results can help to find the landmarks useful for clinical decision making, such as deriving the H-line and acetabular index [15]. We conduct subjective experiments with an orthopedic surgeon with more than 20 years of clinical experience (D1) and a medical imaging practitioner with five years of experience (D2). We request them to score the original image X_{raw}, bone extraction image X_{bone}, and enhancement image X_{boneEnh} using a subjective score of 1 (or 5), meaning that the image provides the lowest (or highest) diagnostic value.

[1] https://www.kaggle.com/datasets/tommyngx/cgmh-pelvisseg.

[2] This paper makes only 150 images public.

[3] https://github.com/wkentaro/labelme.

Table 1 Performance evaluation of the pelvis extraction module from PXRs

	Dice↑	HD95↓	SSIM↑	LPIPS↓	PSNR(dB)↑
U-Net [29]	0.809	54.056	0.565	0.515	12.331
CycleGAN [42]	0.606	59.327	0.616	0.351	13.796
PELE	**0.945**	**34.723**	**0.923**	**0.083**	**21.856**
PELE (w/o Enh)	0.945	34.723	0.827	0.177	16.934

The 'w/o Enh' option means without enhancement, so the Dice and HD95 metrics are the same as the first row. The best result is in bold, and the second best is underlined

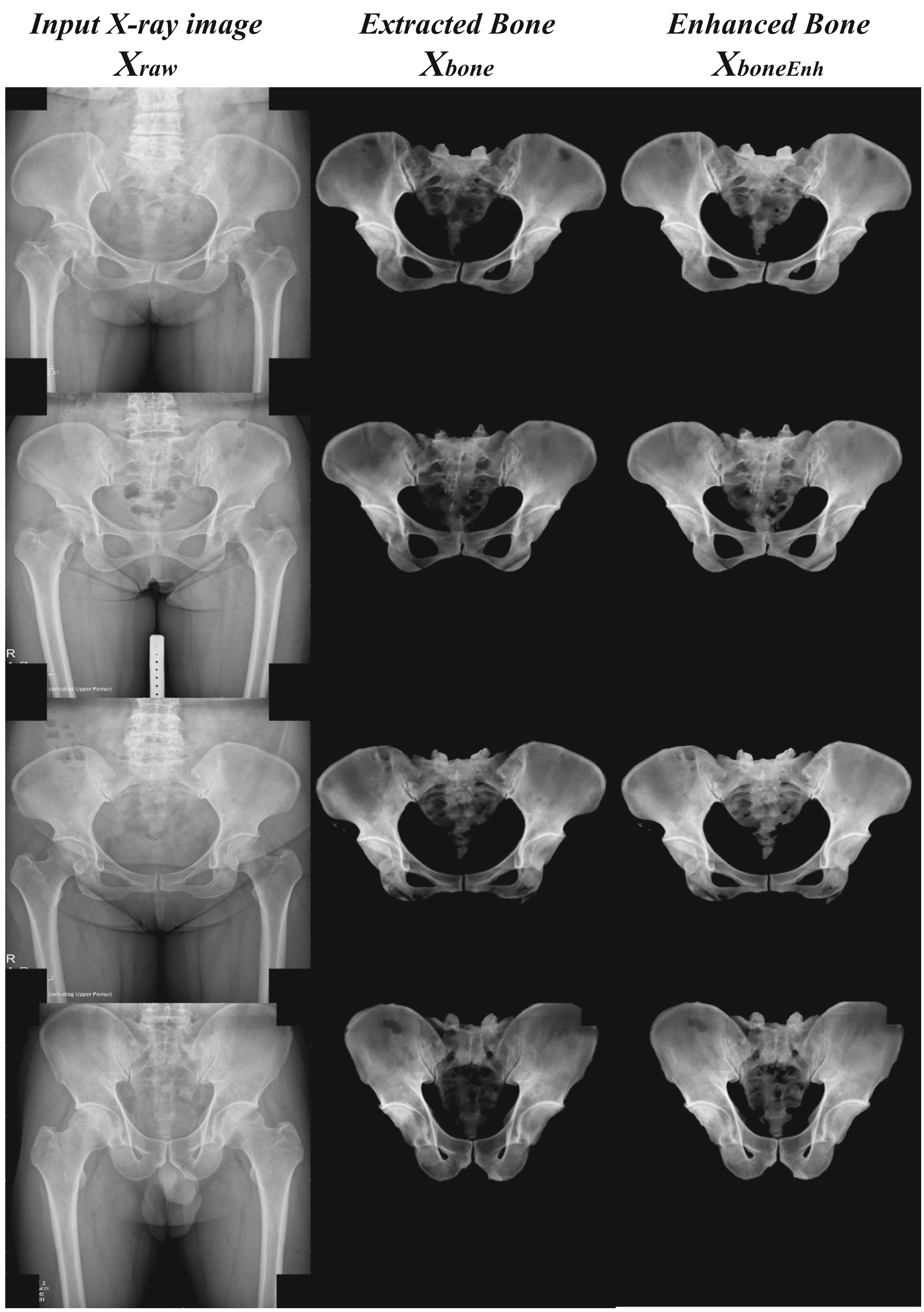

Fig. 4 Visualization of the results obtained by the PELE module. We choose images with different datasets, shapes, and superposition degrees to demonstrate the qualitative effect better. At the same time, we compare the impact before and after using the enhancement network F_{Enh}, proving the F_{Enh} is also indispensable

Images with higher scores are of higher quality and more helpful in locating landmarks needed in clinical decision making.

To reduce the inconsistencies and uncertainties, we do the following settings: the experiment is performed four times over different durations: workday morning, workday evening, weekend morning, and weekend evening. We randomly select 15% of PXRs for testing each time to ensure unbiased selection.

The test results are listed in Table 2: Our method obtains mean scores of 4.23 (D1), 4.31 (D2) (before F_{Enh}), 4.52 (D1), and 4.55 (D2) (after F_{Enh}), Both have significant improvements over X_{raw}.

The performance of landmark detection

We select 14 corresponding landmarks according to CE Angle, acetabular index, H-line, and Perkin quadrant [20, 21, 34], which are commonly used in clinical auxiliary diagnosis, as shown in Fig. 2b. All images are annotated by a pelvic surgeon with over ten years of experience and reexamined by D2, who mainly examines the locations that make up the CE Angle, acetabular index, H-line, etc., and makes corrections, if necessary, after consultation with D1.

For evaluation, we use mean radial error (MRE) to measure the Euclidean distance between ground truth and the predicted result and successful detection rate (SDR) in 4 radii of 3, 4, 6, and 9 pixels(px) in Tables 3 and 4 radii of 2, 2.5, 3, 4 mms(mm) in Table 4. We report Dataset3 separately in mm and compare it with other methods, as shown in Table 4.

Tables 3 and 4 show quantitative performances of different baselines before and after using PELE. To reflect the robustness of the model with various data scales, we train our model using 12.5%(107), 25%(213), and 50%(425) of training data with a simple cross-validation strategy and verify the comprehensive effectiveness of the PELE module. When training with only 107 PXRs (12.5% of all PXRs), the MRE improvement is particularly significant, with a gain of more than 200% over the baseline, demonstrating that PELE is potentially a huge advantage for small datasets. GU2Net does not achieve the best results here, mainly because its training here does not use datasets of multiple organs, so its effect degenerates into an ordinary U-Net.

When trained with 50% data (i.e., 425 PXRs), our method achieves 1.83 MRE and 94.41% SDR within a 9px distance. Also, PELE works well with different baselines (i.e., U-Net, GU2Net), shows good generalization, and can be applied to other baselines. However, our results present a large variation as the standard deviation reaches 9.32px. This is due to the presence of extreme superposition in some PXRs such as those in Fig. 1, and even our experts admit that they have difficulty marking the landmarks well. We also select ten extreme case PXRs and ten normal case PXRs for landmark marking by D1 and D2, respectively. Then we calculate the difference in terms of MRE labeled by two experts on extreme case PXRs and normal case PXRs. The result is that the MRE on normal case PXRs is 1.42px, which is far less than 3.86px on extreme case PXRs, proving the difficulty of labeling extreme case PXRs from an inter-expert perspective.

It is evident that the PELE module brings performance improvement in landmark detection, especially comparing the difference before and after using the PELE module. Furthermore, our model also reaches state-of-the-art results in terms of absolute values. As shown in Table 4, the last four rows present the landmark detection performances of seven competing methods [3, 20, 21, 25]. It is challenging to give an entirely fair comparison as different methods use different protocols and datasets; nevertheless, our MRE of 0.71 mm trained with only 425 PXRs compares favorably to the MRE of 0.93 mm by Liu et al. [20], trained with 8000 PXRs as shown in Table 4.

As shown in Fig. 5, the pink points represent ground truth, and the blue points represent predictions. Obviously, the detected landmarks have been much better localized after PELE is used, in the sense that, after using the PELE module, the distance between the predicted point and the real point becomes smaller visually.

Discussion

In this study, we develop the PELE for extracting pelvis bones in the PXRs and achieve state-of-the-art performance in the landmark detection experiment. Our proposed PELE is specially designed to handle the challenges of pelvis-tissue superposition and is inspired by other machine-learning tasks on other organs like the abdomen and chest, e.g., suppressing ribs to improve lung disease classification or detection [13, 17], enhance the model with realistic simulated images [11] using DeepDRR. We thought of using a decomposition network to separate the soft tissue from the pelvis bone, which is the prototype of our idea. In the design, it was found that merely F_{DE} cannot make good use of the structural knowledge of CT, so the intermediary of DeepDRR is used. CycleGAN [42] is a widely used method for domain adaptation by constructing the mapping of the source-target domains. In the proposed PELE, the structural features of DRR images are adaptively transferred to PXRs using CycleGAN as F_{DA} network. Finally, we design a F_{Enh} module to enhance the image by filtering and tone mapping, aiming at the problem that sometimes contains artifacts in the hip, sacrum, tail, and other parts. Sufficient experiments prove that our method not only achieves the state-of-art result but

Table 2 The score of original image X_{raw}, bone extraction image X_{bone}, and enhancement image X_{boneEnh} by D1 and D2

	Dataset1			Dataset2			Dataset3		
	X_{bone}	X_{boneEnh}	X_{raw}	X_{bone}	X_{boneEnh}	X_{raw}	X_{bone}	X_{boneEnh}	X_{raw}
D1	4.26	4.54	4	4.15	4.50	4	4.28	4.52	4
D2	4.38	4.61	4	4.23	4.45	4	4.33	4.59	4

The score of 1 (or 5) provides the lowest (or highest) diagnostic value. Note that we artificially set the score of X_{raw} to 4 for easier comparison

Table 3 Performances of different landmark detection models before and after using the PELE module

Models	Training Data	MRE (px)	STD (px)	SDR (%) 3px	4px	6px	9px
GU2Net [41]	107	11.64	30.79	52.40	<u>67.34</u>	82.84	90.50
GU2Net with PELE	107	<u>4.85</u>	**12.95**	**53.43**	**67.40**	**83.40**	<u>91.30</u>
U-Net [29]	107	10.70	30.99	51.12	66.13	81.82	90.07
U-Net with PELE	107	**4.74**	<u>14.33</u>	<u>52.79</u>	66.42	<u>83.17</u>	**91.52**
GU2Net [41]	213	6.66	23.44	<u>54.30</u>	<u>70.00</u>	<u>85.97</u>	<u>92.26</u>
GU2Net with PELE	213	**4.39**	<u>11.12</u>	**54.53**	**70.71**	**85.99**	**93.04**
U-Net [29]	213	6.98	22.01	52.30	68.05	83.45	91.30
U-Net with PELE	213	<u>4.72</u>	**10.65**	53.70	69.27	84.51	92.24
GU2Net [41]	425	3.81	20.53	<u>56.54</u>	73.10	87.04	93.35
GU2Net with PELE	425	<u>2.01</u>	<u>9.75</u>	**56.89**	**73.49**	**88.14**	**94.41**
U-Net [29]	425	3.41	19.48	55.17	72.63	86.65	92.70
U-Net with PELE	425	**1.83**	**9.32**	55.49	<u>73.18</u>	<u>87.33</u>	<u>93.38</u>

The best results are in bold, and the second best results are underlined

Table 4 Performances of different landmark detection models before and after using the PELE module

Models	Training Data	MRE (mm)	SDR (%) 2 mm	2.5mm	3 mm	4 mm
GU2Net[41]	425	1.95	84.79	90.56	95.36	96.10
GU2Net with PELE	425	<u>1.03</u>	<u>85.28</u>	**90.83**	**96.21**	**97.13**
U-Net[29]	425	1.32	85.17	90.43	95.07	95.92
U-Net with PELE	425	**0.71**	**85.92**	<u>90.75</u>	<u>96.02</u>	<u>96.74</u>
Liu et al. [25]*	566	2.10	**84.73**	**90.89**	92.57	**96.55**
Liu et al. [21]*	7710	<u>1.24</u>	83.85	–	<u>95.18</u>	–
Liu et al. [20]*	8000	**0.93**	<u>84.29</u>	<u>90.36</u>	**95.83**	–
Benjamin et al. [3]*	122	2.60	–	–	–	–

The difference with Table 3 is that we reports the physical distance (mm) here. The best results are in bold, and the second best results are underlined. –: No experimental results can be found in the original paper. ∗: The performances that are copied from the original paper

also applies to various backbones and can have corresponding improvements.

Ablation Study We provide ablation experiments of the key component modules of our proposed approach: F_{DE}, F_{DA}, and F_{Enh}. We compare the quantitative results with and without these modules, using SSIM and PSNR to evaluate image quality, MRE and SDR to measure landmark detection effectiveness. 425 PXRs are used to train the model, and only 9px distance is selected for SDR. An U-Net is used as the backbone of landmark detection.

The results in Tables 3 and 4 show that when only F_{DE} or F_{DA} module is used to form PELE, the landmark detection performance (e.g., MRE = 5.83px for F_{DE} and MRE = 6.81px for F_{DA}) is even worse than that of directly using the baseline model (e.g., U-Net). This is because the extracted X_{bone} lacks a lot of details, and there are problems such as incomplete soft tissue elimination. However, when F_{DE} and F_{DA} networks

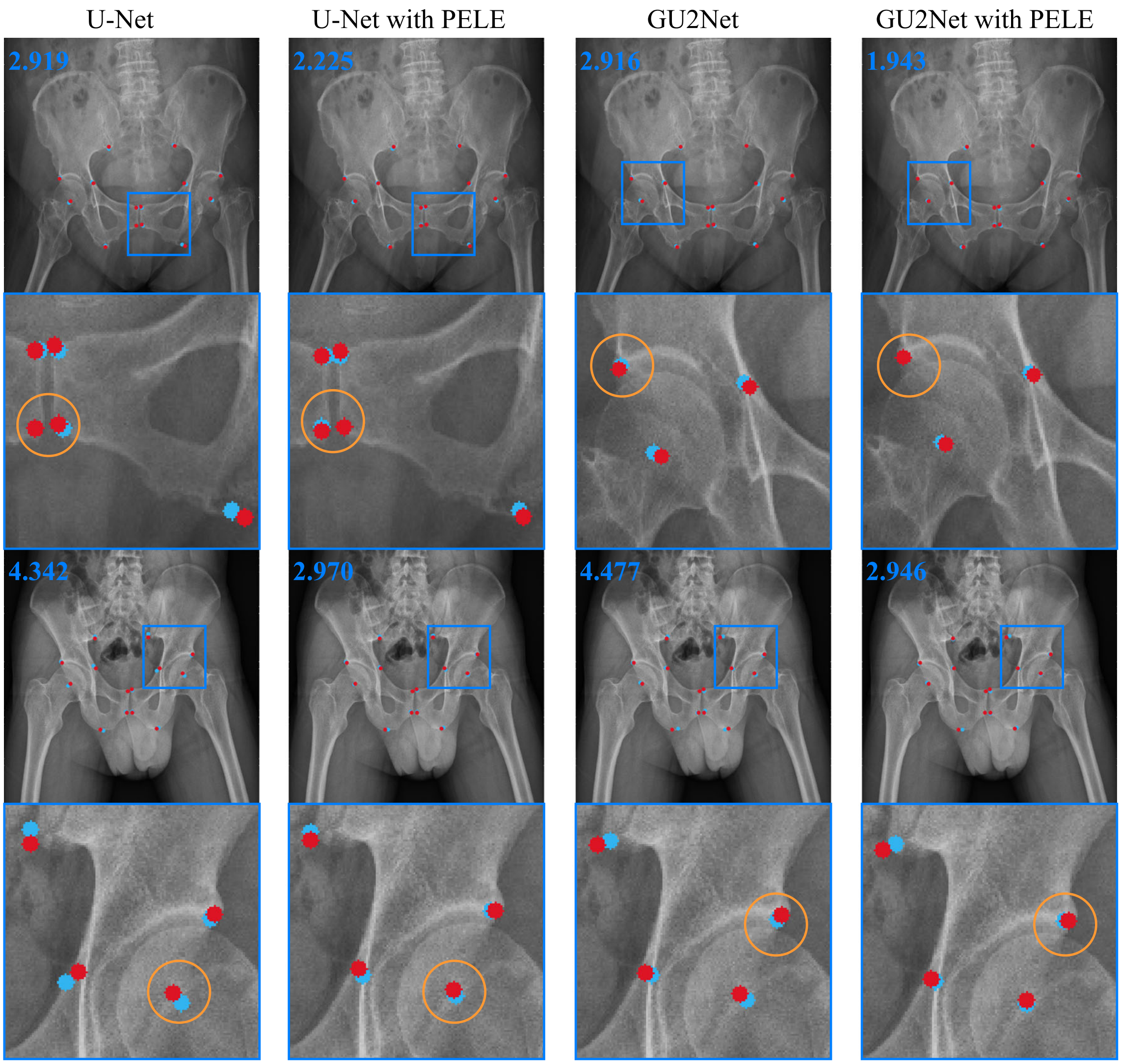

Fig. 5 Visualizations of different approaches under 213 training data setting. The blue points represent the predicted landmarks, while the pink points denote the ground truth labels. The corresponding local details are shown in the following line to demonstrate the results better. The MRE value is displayed on the top left for comparison

are combined, both image quality and landmark detection performances are greatly improved, achieving 2.07px MRE and 93.29% SDR within a 9px distance, which indicates that both F_{DE} and F_{DA} are indispensable. The F_{Enh} mainly enhances the detail information, so the quality indicators such as PSNR and SSIM are improved more greatly and the landmark detection performances reach 1.83px MRE and 93.38% SDR within a 9px distance, which indicate a small improvement (Table 5).

Conclusion

In this paper, we propose a two-stage model, PELE, via learning from the prior structural knowledge of CT to separate the soft tissue and bone from the 2D PXRs and increase the accuracy of landmark detection. In order to better capture the details of the pelvis, we further design a bone enhancement network as post-processing to enhance the details. We perform extensive experiments on multiple PXR datasets, and both qualitative and quantitative results show superiority over the competing methods. We believe that the PELE module, which is suitable for landmark detection, also has a positive

Table 5 Ablation study on different modules

Baseline	F_{DE}	F_{DA}	F_{Enh}	MRE	STD	SDR(9px)	SSIM	PSNR(dB)
✓				3.41	19.48	92.70	–	–
	✓			5.83	19.75	91.08	0.565	12.331
		✓		6.81	20.83	90.87	0.616	13.796
	✓	✓		<u>2.07</u>	<u>13.71</u>	<u>93.29</u>	<u>0.827</u>	<u>16.934</u>
	✓	✓	✓	**1.83**	**9.32**	**93.38**	**0.923**	**21.856**

The best result is in bold, and the second best is underlined. Baseline is a U-Net model which can be referred in Table 3

significance in the diagnosis and treatment of other clinical tasks, which we plan to investigate in future.

Funding This work was supported byNatural Science Foundation of Chinaunder Grant62271465andOpen Fund Project of Guangdong Academy of Medical Sciences, China(No.YKY-KF202206).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The IRB approval is obtained.

Informed consent For the in-house data sets retrospectively collected, there is no informed consent needed.

References

1. Alansary A, Oktay O, Li Y, Le Folgoc L, Hou B, Vaillant G, Kamnitsas K, Vlontzos A, Glocker B, Kainz B, Rueckert D (2019) Evaluating reinforcement learning agents for anatomical landmark detection. Med Image Anal 53:156–164
2. Aubert B, Cresson T, De Guise J, Vazquez C (2022) X-ray to DRR images translation for efficient multiple objects similarity measures in deformable model 3D/2D registration. IEEE Trans Med Imaging 42(4):897–909
3. Aubert B, Vazquez C, Cresson T, Parent S, De Guise J (2016) Automatic spine and pelvis detection in frontal X-rays using deep neural networks for patch displacement learning. In: 2016 IEEE 13th international symposium on biomedical imaging (ISBI). IEEE, pp 1426–1429
4. Avisdris N, Joskowicz L, Dromey B, David AL, Peebles DM, Stoyanov D, Ben Bashat D, Bano S (2022) Biometrynet: landmark-based fetal biometry estimation from standard ultrasound planes. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 279–289
5. Barbu A, Suehling M, Xu X, Liu D, Zhou SK, Comaniciu D (2010) Automatic detection and segmentation of axillary lymph nodes. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 28–36
6. Bier B, Unberath M, Zaech JN, Fotouhi J, Armand M, Osgood G, Navab N, Maier A (2018) X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 55–63
7. Browning J, Kornreich M, Chow A, Pawar J, Zhang L, Herzog R, Odry BL (2021) Uncertainty aware deep reinforcement learning for anatomical landmark detection in medical images. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp 636–644
8. Cheng CT, Wang Y, Chen HW, Hsiao PM, Yeh CN, Hsieh CH, Miao S, Xiao J, Liao CH, Lu L (2021) A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun 12(1):1066
9. Duan J, Bello G, Schlemper J, Bai W, Dawes TJ, Biffi C, de Marvao A, Doumoud G, O'Regan DP, Rueckert D (2019) Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. IEEE Trans Med Imaging 38(9):2151–2164
10. Elkhill C, LeBeau S, French B, Porras AR (2022) Graph convolutional network with probabilistic spatial regression: application to craniofacial landmark detection from 3d photogrammetry. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 574–583
11. Gao C, Killeen BD, Hu Y, Grupp RB, Taylor RH, Armand M, Unberath M (2023) Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nat Mach Intell 5(3):294–308
12. Ge J, Saeidi H, Opfermann JD, Joshi AS, Krieger A (2019) Landmark-guided deformable image registration for supervised autonomous robotic tumor resection. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer, pp 320–328
13. Han L, Lyu Y, Peng C, Zhou SK (2022) Gan-based disentanglement learning for chest X-ray rib suppression. Med Image Anal 77:102369
14. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18(2):203–211
15. Karnik A, Lawande A, Lawande MA, Patkar D, Aroojis A, Bhatnagar N (2021) Practice essentials of imaging in early diagnosis of DDH. Indian J Orthop 55:1–14
16. Kasten Y, Doktofsky D, Kovler I (2020) End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images. In: Machine learning for medical image reconstruction: third international workshop, MLMIR 2020, held in conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3. Springer, pp 123–133

17. Li H, Han H, Li Z, Wang L, Wu Z, Lu J, Zhou SK (2020) High-resolution chest X-ray bone suppression using unpaired CT structural priors. IEEE Trans Med Imaging 39(10):3053–3063
18. Li W, Lu Y, Zheng K, Liao H, Lin C, Luo J, Cheng CT, Xiao J, Lu L, Kuo CF, Miao S (2020) Structured landmark detection via topology-adapting deep graph learning. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer, pp 266–283
19. Li Y, Alansary A, Cerrolaza JJ, Khanal B, Sinclair M, Matthew J, Gupta C, Knight C, Kainz B, Rueckert D (2018) Fast multiple landmark localisation using a patch-based iterative network. In: Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. Springer, pp 563–571
20. Liu C, Xie H, Zhang S, Mao Z, Zhang Y (2020) Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip. IEEE Trans Med Imaging 39(99):1–1
21. Liu C, Xie H, Zhang S, Xu J, Sun J, Zhang Y (2019) Misshapen pelvis landmark detection by spatial local correlation mining for diagnosing developmental dysplasia of the hip. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. Springer, pp 441–449
22. Liu D, Zhou SK, Bernhardt D, Comaniciu D (2010) Search strategies for multiple landmark detection by submodular maximization. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2831–2838
23. Liu J, Xing F, Shaikh A, Linguraru MG, Porras AR (2022) Learning with context encoding for single-stage cranial bone labeling and landmark localization. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 286–296
24. Liu P, Han H, Du Y, Zhu H, Li Y, Gu F, Xiao H, Li J, Zhao C, Xiao L, Wu X, Zhou S (2021) Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. Int J Comput Assist Radiol Surg 16:749–756
25. Liu W, Wang Y, Jiang T, Chi Y, Zhang L, Hua XS (2020) Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. In: Medical image computing and computer assisted intervention–MICCAI 2020: 23rd international conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer, pp 670–679
26. Mader AO, von Berg J, Fabritz A, Lorenz C, Meyer C (2018) Localization and labeling of posterior ribs in chest radiographs using a CRF-regularized FCN with local refinement. In: Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11. Springer, pp 562–570
27. Noothout JM, De Vos BD, Wolterink JM, Postma EM, Smeets PA, Takx RA, Leiner T, Viergever MA, Išgum I (2020) Deep learning-based regression and classification for automatic landmark localization in medical images. IEEE Trans Med Imaging 39(12):4011–4022
28. Payer C, Štern D, Bischof H, Urschler M (2016) Regressing heatmaps for multiple landmark localization using CNNs. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 230–238
29. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer, pp 234–241
30. Ruiz Santiago F, Santiago Chinchilla A, Ansari A, Guzmán Álvarez L, Castellano García MdM, Martínez Martínez A, Tercedor Sánchez J (2016) Imaging of hip pain: from radiography to cross-sectional imaging techniques. Radiol Res Pract **2016**, 6369237
31. Sofka M, Wetzl J, Birkbeck N, Zhang J, Kohlberger T, Kaftan J, Declerck J, Zhou SK (2011) Multi-stage learning for robust lung segmentation in challenging CT volumes. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 667–674
32. Unberath M, Zaech JN, Lee SC, Bier B, Fotouhi J, Armand M, Navab N (2018) Deepdrr—a catalyst for machine learning in fluoroscopy-guided procedures. In: Medical image computing and computer assisted intervention–MICCAI 2018: 21st international conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11. Springer, pp 98–106
33. Wang CW, Huang CT, Hsieh MC, Li CH, Chang SW, Li WC, Vandaele R, Marée R, Jodogne S, Geurts P, Chen C, Zheng G, Chu C, Mirzaalian H, Hamarneh G, Vrtovec T, Bulat I (2015) Evaluation and comparison of anatomical landmark detection methods for cephalometric X-ray images: a grand challenge. IEEE Trans Med Imaging 34(9):1890–1900
34. Wang Y, Lu L, Cheng CT, Jin D, Harrison AP, Xiao J, Liao CH, Miao S (2019) Weakly supervised universal fracture detection in pelvic X-rays. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. Springer, pp 459–467
35. Xu J, Xie H, Liu C, Yang F, Zhang S, Chen X, Zhang Y (2021) Hip landmark detection with dependency mining in ultrasound image. IEEE Trans Med Imaging 40(12):3762–3774
36. Yao Q, Quan Q, Xiao L, Kevin Zhou S (2021) One-shot medical landmark detection. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. Springer, pp 177–188
37. Zhang J, Liu M, Shen D (2017) Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans Image Process 26(10):4753–4764
38. Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen SGF, Tang Z, Chen KC, Xia JJ, Shen D (2020) Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. Med Image Anal 60:101621
39. Zhong Z, Li J, Zhang Z, Jiao Z, Gao X (2019) An attention-guided deep regression model for landmark detection in cephalograms. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. Springer, pp 540–548
40. Zhou XY, Lai B, Li W, Wang Y, Zheng K, Wang F, Lin C, Lu L, Huang L, Han M, Xie G, Xiao J, Kuo Cf, Harrison A, Miao S (2021) Scalable semi-supervised landmark localization for X-ray images using few-shot deep adaptive graph. In: Deep generative models, and data augmentation, labelling, and imperfections: first workshop, DGM4MICCAI 2021, and first workshop, DALI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1. Springer, pp 145–153
41. Zhu H, Yao Q, Xiao L, Zhou SK (2021) You only learn once: universal anatomical landmark detection. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer, pp 85–95
42. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.