IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 November 2023, accepted 27 November 2023, date of publication 30 November 2023,
date of current version 8 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3337824

RESEARCH ARTICLE

Multi-Scale Semantic Fusion of a Large Receptive Field for Irregular Pelvic X-Ray Landmark Detection

CHENYANG LU[1], JIAGENG ZHAO[2], WEI CHEN[3], (Member, IEEE), XU QIAO[1], (Member, IEEE), AND QINGYUN ZENG[2]

[1]School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China
[2]Massage Department, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, Shandong 250061, China
[3]Department of Radiology, Shandong First Medical University, Taian, Shandong 271000, China

Corresponding authors: Xu Qiao (qiaoxu@sdu.edu.cn) and Qingyun Zeng (qy_zeng2000@126.com)

This work was supported in part by the Shandong Province Natural Science Foundation under Grant ZR202103030517, and in part by the National Natural Science Foundation of China under Grant U1806202 and Grant 82071148.

ABSTRACT Pelvic landmark detection is a significant pre-task to measure the clinical measurement in pelvic abnormality analysis. Accurate pelvic landmark detection could provide reliable clinical parameter measurement results, which are helpful for doctors to diagnose and treat pelvic diseases. However, the multi-scale characteristics, temporal diversity, and pathological abnormalities of different pelvic X-rays bring enormous challenges to the landmark detection task. In order to retain strong robustness in irregular pelvic X-rays, we propose a novel, flexible two-stage framework. In the initial stage, a single neural network is employed to estimate the locations of every landmark simultaneously, enabling the identification of potential landmark regions. Then, the receptive field of candidate region proposals is expanded by 4 times through the receptive field amplification module. In the second stage, the landmark detection module fuses semantically rich features at different scales through a multi-scale semantic fusion module. So that the framework can fully learn the strongly relevant semantic information around the landmark at high resolution. We collected a data set of 430 pelvic X-rays, including a large number of irregular pelvic X-rays, to evaluate our framework. The experimental results demonstrate that our framework achieves a state-of-the-art detection mean radial error of 3.724 $\pm$ 4.247-mm. The experimental results show that the proposed method can help doctors quickly and accurately find the characteristic points of the pelvis and could be applied to clinical diagnosis.

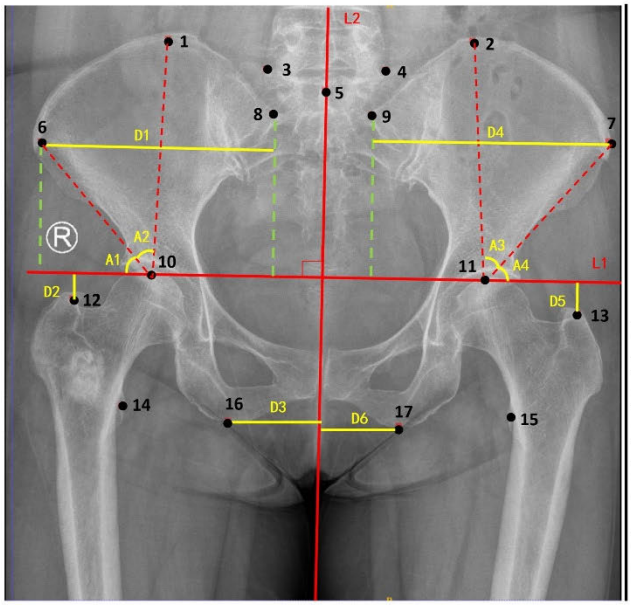
INDEX TERMS Landmark detection, irregular pelvic X-ray, receptive field amplification, multi-scale semantic fusion.

I. INTRODUCTION

Accurate and reliable detection of anatomical landmarks is a crucial preprocessing step for therapy planning and intervention in various medical scenarios [1], including knee joint surgery [2], bone age estimation [3], carotid artery bifurcation [4], and vertebral trauma surgery [5]. Also, it plays an important role in medical image analysis, such as the initialization of registration [6] or segmentation algorithms [7]. However, up to now, far too little practice of machine learning has been carried out on the pelvis. Accurate measurement of clinical measurements of the pelvis are help reveal pelvic abnormalities.

In clinical practice, pelvic abnormality analysis is usually done manually. Pelvic landmark detection is the pre-task of pelvic abnormality analysis, and accurate landmark detection is the premise of reliable clinical measurements. Figure 1 illustrates the spatial arrangement of the 17 anatomical landmarks alongside the corresponding clinical measurements of the pelvis. Four clinically measured angle values, denoted as A1 to A4, are commonly employed to assess the degree of pelvic anteversion and tilting, while six clinically measured

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



L1: Highest point of the right iliac crest
L2: Highest point of the left iliac crest
L3: Midpoint of the paraspinal sulcus of the right superior sacrum
L4: Midpoint of the paraspinal sulcus of the left superior sacrum
L5 : Midpoint of the first sacrum
L6: Right anterior superior iliac spine
L7: Left anterior superior iliac spine
L8: Right posterior superior iliac spine
L9: Left posterior superior iliac spine
L10: Highest point of right femur
L11: Highest point of left femur
L12: Greater trochanter of the right femur
L13: Greater trochanter of the left femur
L14: Lesser trochanter of the right femur
L15: Lesser trochanter of the left femur
L16: Right ischial tuberosity
L17: Left ischial tuberosity

Clinical measurements:
A1: Right iliac angle
A2: Right CE angle
A3: Left CE angle
A4: Left iliac angle
D1: Right iliac wing width
D2: Right hip joint space
D3: Sciatic tubercle distance 3
D4: Left Iliac wing width
D5: Left hip joint space
D6: Sciatic tubercle distance 6

FIGURE 1. Pelvic landmarks and pelvic clinical measurements in a pelycogram.

distance values, designated as D1 to D6, are typically used to measure parameters such as iliac wing width. The manual tracing of pelvic landmarks on radiographs is a tedious and time-consuming process.

Even experienced clinicians may require 20-30 minutes to perform a single pelvic X-ray analysis [8]. Additionally, there is a significant risk of inter- and intra-observer variabilities, as the accuracy of landmark identification depends on the expertise and experience of the clinician performing the analysis [9]. So clinical pelvic evaluations and resulting treatment decisions are susceptible to the precise estimation of landmark locations. Meanwhile, failure to detect pelvic deformities, and injuries, or to provide timely treatment can result in serious adverse outcomes and costly treatment expenses in the future [5]. Therefore, it is imperative to develop an automatic pelvic landmark detection system that can identify pelvic landmarks accurately, reliably, and rapidly.

Over the past few decades, numerous automatic anatomical landmark detection methods based on machine learning have been proposed, including rule-based methods [10], template-matching methods [11], and active appearance model methods [12]. These methods were designed to identify anatomical landmarks. Subsequently, neural networks, support vector machines, and random forests have been used for landmark localization [13]. However, these methods can't achieve the high precision of detection demanded by clinical practice.

In recent years, deep learning has gained widespread use in medical image analysis, including anatomical landmark detection [14], [15], [16], [17]. The published deep learning landmark detection frameworks can be broadly divided into two categories: one is the end-to-end framework [16], [18], [19], [20]. This direct regression method is limited by the image size that the framework can handle. The image often needs to be reduced to 3-4 times the original for reprocessing. However, this practice compromises the intrinsic high-resolution attributes of medical images, thereby impeding the model's ability to comprehensively capture contextual semantic information. The other is the multi-stage framework [21], [22], [23], scholarly reviews have established that the multi-stage framework generally outperforms the end-to-end framework in terms of overall detection accuracy [24], [25]. For these frameworks, firstly, the coarse positions of all landmarks are estimated, identifying potential candidate regions likely to contain the target landmarks. In the second stage, the candidate regions are extracted and subsequently employed for training to enhance the precision of the landmarks roughly detected in the initial stage. However, the first stage of these methods is fixed structure and only rough detection of landmarks to extract candidate region proposals from a small receptive field. Most importantly, these methods simply train the candidate region proposals of a small receptive field in the second stage, which will lead to an inability to fully learn the strongly relevant semantic information around landmarks and the individual and anatomical differences of the pelvis at multiple scales. Which is not suitable for the clinical pelvic X-ray with time diversity and pathological diversity. The ground truth may suffer from issues regarding generalizability and reliability. Owing to disparities between the imaging process and the sampled data, the resultant candidate region proposals exhibit variations in their respective receptive fields. This phenomenon is illustrated through the highlighted red boxes in Figure 2. As depicted in the figure, pelvic marker detection encounters a significant challenge arising from the temporal misalignment between pelvic X-ray images and

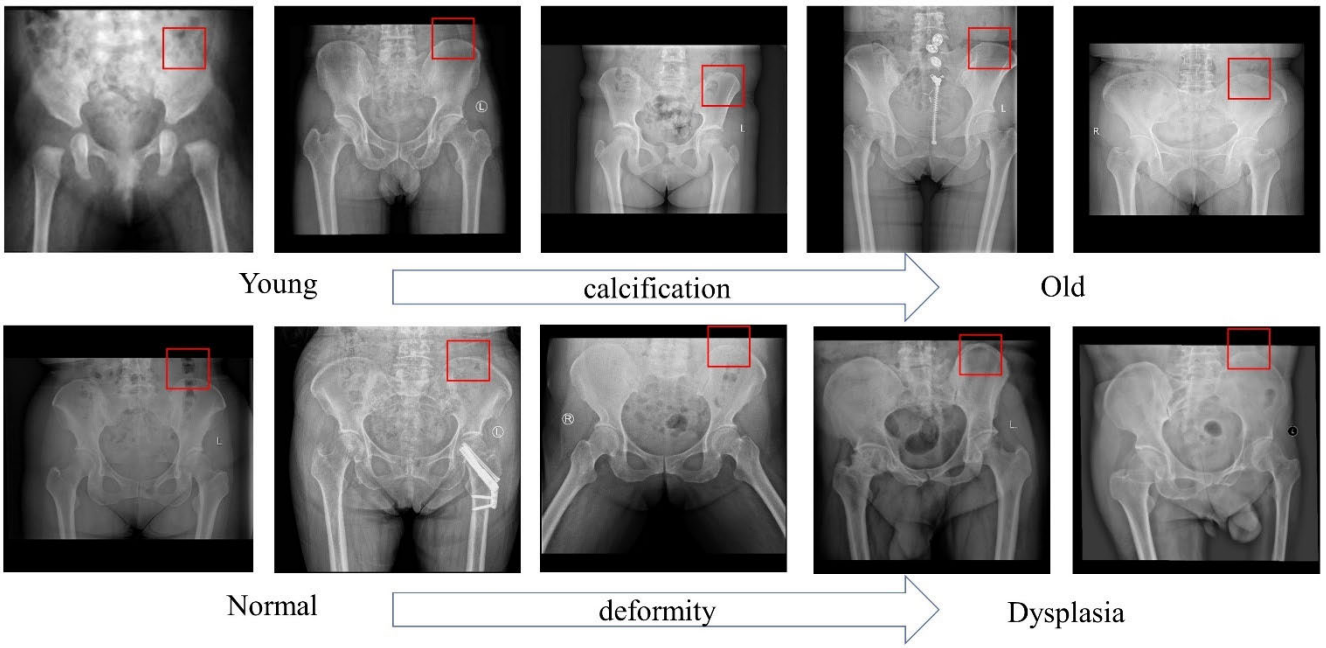


FIGURE 2. The first line shows the challenge of time diversity, at different stages of bone calcification, the morphology of the markers showed diversity. The second line is the challenge of pathological physical distortion, different degrees of deformation lead to different signs of wear and deformity.

pathological abnormalities. Temporal variation is evident in the morphological characteristics of these markers, which exhibit diversity during various stages of bone calcification. This diversity is observable through the blurring of calcified regions and the emergence of bone spurs in proximity to the ischial tubercle nodes.

Furthermore, pathological diversity introduces variations in pelvic structure, resulting in varying degrees of deformation and displacement. These factors collectively pose obstacles for computer models seeking to learn from these data. Thus, pelvic landmark detection is a complex and critical challenge for computers, requiring time and experience-sensitive.

In order to solve the appeal problem, we propose a novel, flexible two-stage framework. Our framework offers the flexibility to use various network architectures as a backbone without any constraints. Most importantly, in order to bridge the semantic gap between features extracted from complex candidate region proposals, we adopt a multi-scale semantic fusion module (MSFM), which integrates features with high resolution and weak semantics with features with low resolution and strong semantics. MSFM extracts features with different resolutions from candidate region proposals and re-weights them so that the model pays more attention to the information near landmarks after fusing multi-scale features. Not by modifying the backbone network [26] or federated learning [27]. Moreover, we introduce a receptive field amplification module (RFAM), designed to expand the receptive field of the landmark detection module by approximately fourfold. This augmentation enables the model to comprehensively assimilate pertinent semantic information in the vicinity of landmarks. Notably, this enables the framework to learn more about global and local semantic information. Furthermore, beyond the assessment of detection accuracy, we have undertaken an evaluation of clinical measurements pertinent to pelvic analysis for the first time.

The contributions of this paper are as follows:

- We have introduced a novel flexible two-stage framework that, for the first time, addresses the challenge of landmark detection in irregular pelvic X-rays while simultaneously evaluating the pertinent clinical measurements of the pelvis.
- We introduce a MSFM, designed to generate semantically enriched features through the amalgamation of multi-scale candidate region proposal feature mappings. Additionally, we present the RFAM, which expands the receptive field of candidate region proposals by a factor of nearly four, facilitating the comprehensive acquisition of strongly correlated semantic information in the vicinity of landmarks within our framework.
- Experiments proved the superiority of our framework when compared to state-of-the-art methods. Our framework could be an efficient and accurate landmark detection method for doctors to do pelvic abnormality analysis.

The structure of the remainder of this paper is as follows: Section II reviews related work in the field. Section III outlines the construction method of the proposed model, followed by the experiments and results presented in Section IV. Section V discusses the observations and findings derived

from the experiments. Finally, Section VI provides concluding remarks and suggestions for future work.

II. RELATED WORK

A. PELVIC CLINICAL MEASUREMENTS ANALYSIS

As far as we know, most of the published literature [28], [29] mainly focuses on the development of infant hip bones. The datasets used in these studies predominantly consist of standardized X-ray images of the infant's pelvis, and the scope of landmark detection is limited to a few points around the hip bone. Consequently, the inherent simplicity of this task is evident. Limited research endeavors have been directed towards the detection of a substantial quantity of landmarks across the pelvis. Statchen et al. [30] trained a U-net [31] with 902 pelvic X-rays. The objective was to detect 22 distinct landmarks on these X-ray images and investigate the potential utility of these automated measurements for predicting femoral fractures within a machine learning framework. Bier et al. [5] present a method to automatically detect anatomical landmarks in X-ray images independent of the viewing direction. This method successfully detected 23 landmarks and realized X-ray pose estimation together with preoperative CT. The achieved detection accuracy was 5.6 ± 4.5mm. Zhu et al. [32] developed a universal anatomical landmark detection model that learned once from multiple data sets corresponding to different anatomical regions. The model consists of a local network and a global network, which capture local features and global features, respectively. It achieved a mean radial error of 6.183 ± 19.711mm on the internal pelvic data set. Lu et al. [33] proposed a deep neural network system with prior knowledge of the active shape model (ASM), which was used to automatically detect the landmarks on the pelvis. The pelvic contour was extracted through the ASM, enabling a rough detection of landmarks. Subsequently, leveraging the acquired prior knowledge, the deep neural network facilitated precise landmark detection. The efficacy of the system was empirically assessed, revealing a mean radial error of 4.159 ± 5.015 mm.

Unfortunately, these aforementioned investigations failed to undertake an analysis of pelvic clinical measurements. Furthermore, each study independently employed distinct internal datasets, thereby impeding the comparison of methodologies. In our subsequent experimentation, we adopted the approach in the aforementioned study to train and evaluate our irregular datasets. The outcomes of these experiments indicate that the accuracy achieved by the methods discussed above is less than desirable. In contrast, the framework we present in this paper shows the highest accuracy.

B. LANDMARK DETECTION IN MEDICAL IMAGE

The integration of artificial intelligence brought about a paradigm shift in the healthcare industry. In response, researchers have turned to the cutting-edge capabilities of deep learning to propel landmark detection research and practice to new heights. Qian et al. [34] introduced CephaNet, the first faster RCNN-based method, which utilizes a multitask loss to reduce intra-class variations and a two-stage repair strategy to eliminate superfluous or undetected landmarks. Similarly, Lee et al. [35] proposed a two-stage approach that initially extracts potential regions of interest (ROI) for every landmark and subsequently employs a set of Bayesian Convolutional Neural Networks (CNN) to estimate the precise landmark location within the extracted region. Lee et al. [36] proposed a single-channel convolutional neural network to perform accurate landmark detection hierarchically, and the proposed patch-wise method significantly enhanced the local feature encoder, thus further improving the final accuracy. Zeng et al. [37] treated cephalometric landmark detection as a multi-stage regression problem and designed a cascaded three-stage CNN structure with a coarse-to-fine detection strategy. Kwon et al. [38] developed a multi-stage probabilistic approach that simultaneously used local appearances and global features. The method involves the use of a single network for the initial detection of all landmarks, followed by individual refinement of each landmark using high-resolution cropped images, with each refinement step performed by a separate CNN model dedicated to that landmark. Ao et al. [39] proposed a feature aggregation and refinement network (FARNet), which includes a multi-scale feature aggregation module for multi-scale feature fusion and a feature refinement module for high-resolution heat map regression. It was evaluated on three public anatomical landmark detection data sets and achieved the most advanced performance.

The strides made by deep learning-based approaches in medical image landmark detection are commendable. However, so far, in the task of automatic landmark detection, no framework has been proposed for irregular images. In addition, the most advanced methods, whether end-to-end or multi-stage frameworks, still have significant limitations that hinder their application in clinical settings. One of the main limitations is that, while maintaining the high-resolution features of complex medical images, it is difficult to learn features of different scales as much as possible with a large receptive field. Therefore, demand arises for a framework that can handle irregular medical images. This framework should be equipped not only with an expansive receptive field to facilitate comprehensive learning of pertinent semantic information but also with the capability to integrate multi-scale semantic information.

III. METHOD

A. PELVIC CLINICAL MEASUREMENTS ANALYSIS

Define X as the X-ray sample space and Y as the landmark coordinate space. The data set is defined as $D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq X \times Y$. Our task is to train a detector $f_d(\cdot)$ on D to make $y = f_d(x)$. Where x represents sample images, y represents landmark coordinates of samples, n represents the number of samples, and $f_d(\cdot)$ represents the landmark detector.

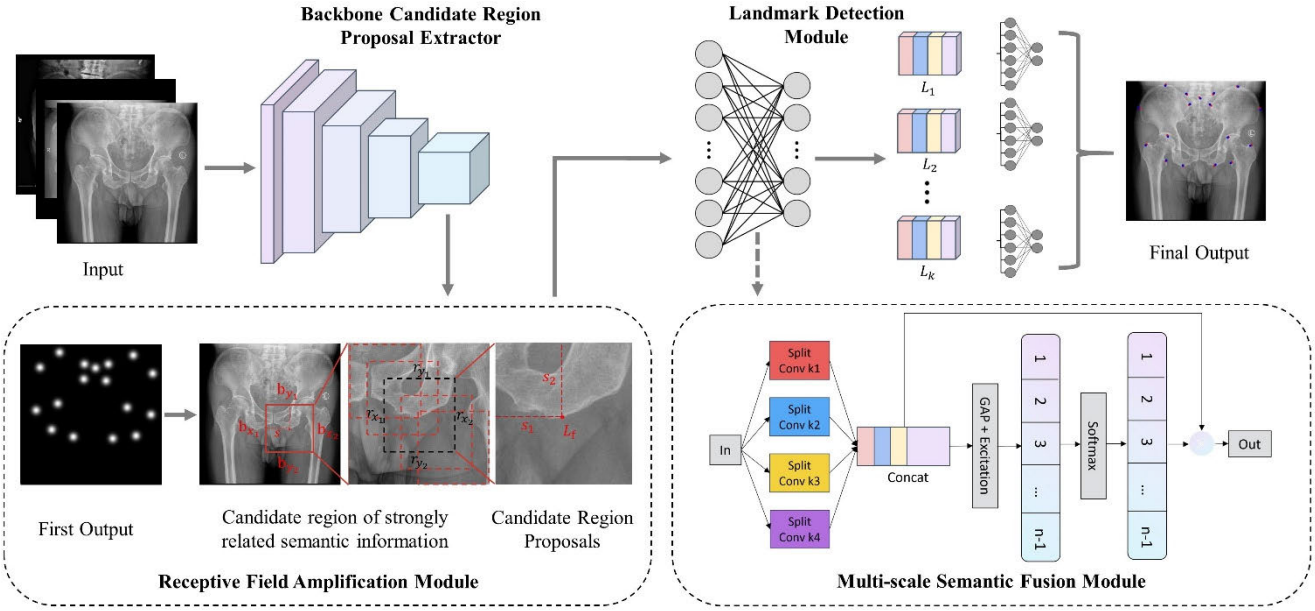


FIGURE 3. Schematic representation of the proposed framework for landmark detection in pelvic X-rays. Initially, the samples get the first rough prediction through the backbone candidate region proposal extractor. Subsequently, the receptive field is augmented through the RFAM. The results of RFAM are input to the landmark detection module for detailed training, and the final prediction is obtained. In which each 3 × 3 convolution in the landmark detection module is replaced by MSFM. Which can fuse semantic information on different scales and make maximum use of limited image resources to learn strongly relevant semantic information.

Note that variations in size and complexity of X-ray samples from different patients lead to deviations in the marginal distribution $P(x)$ of the data, which means $P(x_1) \neq P(x_2) \neq \ldots \neq P(x_n)$.

In our method, we use a multi-scale semantic fusion module $\varphi(\cdot)$ to map $x_1, x_2, \ldots, x_n$ into the high dimensional feature subspace. Q. Make $q_1 = \varphi(x_1), q_2 = \varphi(x_2), \ldots, q_n = \varphi(x_n)$ to get $P(q_1) \approx P(zq_2) \approx \ldots \approx P(q_n)$. Further to obtain $P(\mathrm{y_1}|\mathrm{q_1}) \approx P(\mathrm{y_2}|\mathrm{q_2}) \approx \ldots \approx P(\mathrm{y_n}|\mathrm{q_n})$. Through this process, we could get a better detector $f_d(\cdot)$ to make $\mathrm{y} = f_d(x)$. Among them, the architecture behind the backbone network is used to refine the location of the detected landmarks.

B. FRAMEWORK ARCHITECTURE

As shown in Figure 3, the framework proposed in this paper mainly includes the backbone candidate region proposal extractor, RFAM, landmark detection module, and MSFM.

1) BACKBONE CANDIDATE REGION PROPOSAL EXTRACTOR

The proposed framework uses a convolutional neural network $f(\cdot)$ as candidate a region proposal extractor to convert the input X-ray image into a heatmap h_k of landmarks. Our framework provides the flexibility to use various network architectures as the backbone without any restrictions, such as Resnet50, HRNet, U-net, and so on. This will be explained in detail in the ablation experiment section. Specifically, after the sample image is resized to 512 × 512, the thermal map of the real landmark can be returned through the backbone network, and the predicted rough coordinates can be obtained. We define the backbone candidate region proposal extractor as:

$$f : X \rightarrow h_1, h_2, \ldots, h_k \tag{1}$$

The value at a certain point in the heatmap is

$$h(x, y) = e^{-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma^2}} \tag{2}$$

where x and y are the coordinate values of the point, μ_{x} and μ_{y} are the real marks of the point, and σ is the adjustment parameter.

In this process, in order to get the candidate region proposals in RFAM, we need to ensure that the actual error (converted into actual size) of each landmark in each picture is less than s/2 px as much as possible.

2) RECEPTIVE FIELD AMPLIFICATION MODULE

To fully learn the strongly relevant semantic information of landmarks in the original high-resolution image within an expansive receptive field, we propose RFAM. Our primary concept posits that the correlation between a landmark and its corresponding anatomical structure increases proportionally with the landmark's proximity (Euclidean distance) to said structure. While the entirety of the image's semantic information contributes to landmark detection, the diminished weight of semantic information distal from the point necessitates our strategy of directing the model's learning efforts towards strongly pertinent semantic information proximate to the landmark. Nonetheless, the traditional receptive field for learning local semantic information is too small to fully learn useful features. Therefore, we propose RFAM to fully

learn the strongly relevant semantic information within a large receptive field.

The main workflow of RFAM is shown in Figure 3. First, it is suggested to generate a candidate region of strongly correlated semantic information according to the coordinates predicted by the backbone candidate region proposal extractor, and the method is as follows:

$$\begin{aligned} \mathbf{b}_{x_1} &= \mathbf{L}_f^{(x)} - s;\ \mathbf{b}_{y_1} = \mathbf{L}_f^{(y)} - s \\ \mathbf{b}_{x_2} &= \mathbf{L}_f^{(x)} + s;\ \mathbf{b}_{y_2} = \mathbf{L}_f^{(y)} + s \end{aligned} \tag{3}$$

Here, $\mathbf{L}_f^{(x)}$ and $\mathbf{L}_f^{(y)}$ represent the predicted landmark coordinates along the x and y axes, respectively, $\mathbf{b}_{x_1}$, $\mathbf{b}_{y_1}$, $\mathbf{b}_{x_2}$, and $\mathbf{b}_{y_2}$ represent the four edges of the candidate region proposal, while s denotes the size of the region proposal. This expansive region proposal encapsulates what we perceive as a region of strongly pertinent semantic information pertaining to the landmark. However, the pixel dimensions of this proposed region remain relatively large. Consequently, within this extended regional proposal, n candidate region proposals are randomly generated for utilization during the landmark detection phase. The generation method is as follows:

$$\begin{aligned} \mathbf{r}_{x_1} &= \mathbf{L}_f^{(x)} + s_1;\ \mathbf{r}_{x_2} = \mathbf{L}_f^{(x)} + s_1 + \left(\tfrac{s}{2}\right) \\ \mathbf{r}_{y_1} &= \mathbf{L}_f^{(y)} + s_2;\ \mathbf{r}_{y_2} = \mathbf{L}_f^{(y)} + s_2 + \left(\tfrac{s}{2}\right) \end{aligned} \quad s_1, s_2 \in \mathbb{R} \tag{4}$$

Among them, $\mathbf{r}_{x_1}$, $\mathbf{r}_{y_1}$, $\mathbf{r}_{x_2}$, and $\mathbf{r}_{y_2}$ respectively represent the four edges of the candidate region proposals, while s_1 and s_2 denote the size of the random disturbance values of width and height, respectively. Attention, $s_1, s_2 \in (-s/2, s/2)$. Through this approach, we extract a set of n candidate region proposals for each landmark. Concurrently, the resultant theoretical receptive field expands to approximately four times its original size, thus enabling a more comprehensive acquisition of strongly pertinent information within the vicinity of the landmark. We designate the upper left corner of the image as the origin point. To guarantee the availability of ample candidate region proposals, we pad the periphery of landmarks situated along edges and corners with pixels having a value of 0. The resulting coordinates for the processed landmarks are as follows:

$$\mathbf{L}_f^{(x)} = \begin{cases} s & \text{if } \mathbf{L}_f^{(x)} < s \\ 2\mathbf{L}_f^{(x)} + s - W & \text{if } W - \mathbf{L}_f^{(x)} < s \\ \mathbf{L}_f^{(x)} & \text{else} \end{cases} \tag{5}$$

$$\mathbf{L}_f^{(y)} = \begin{cases} s & \text{if } \mathbf{L}_f^{(y)} < s \\ 2\mathbf{L}_f^{(y)} + s - H & \text{if } H - \mathbf{L}_f^{(y)} < s \\ \mathbf{L}_f^{(y)} & \text{else} \end{cases} \tag{6}$$

where W and H represent the image resolution. After the module outputs the candidate area proposals, during the landmark detection module, the actual landmark positions within these proposals, which are output together, are used as labels. For each landmark, the coverage area of all the candidate areas we obtained is about four times that of the original candidate area, so the receptive field of the landmark detection module is expanded by about four times. Based on prior empirical findings, we establish s as 512 and n as 300. It is imperative to underscore that this segment solely pertains to the training process, with the testing phase necessitating only one candidate region proposal for precise detection.

3) LANDMARK DETECTION MODULE

The landmark detection module is a crucial component of our framework and is designed to detect all anatomical landmarks in an input X-ray image simultaneously. By considering each landmark separately, the module effectively captures the strong relevant semantic information around the landmark and deeply learns the anatomical structure knowledge and semantic information of the landmark. To achieve this, we use candidate region proposals r_k from the *kth* landmark of RFAM as the input of its landmark detection network $g_k(\cdot)$. Among them, the backbone network of the landmark detection module is also unlimited. Our framework provides the flexibility to use various network architectures as the backbone without any restrictions, such as Resnet50, HRNet, U-net, and so on. The output of $g_k(\cdot)$ is a vector $\mathbf{L}g_k \in \mathbb{R}^2$, which represents the predicted x- and y- coordinates of the landmark. The overall module is as follows:

$$\mathbf{L}g_k = g_k\,(r_k) = [\widehat{x_k}, \widehat{y_k}] \tag{7}$$

Here, $\widehat{x_k}$ and $\widehat{y_k}$ represent the predicted x- and y- coordinates of the *kth* landmark, respectively. It is of significance to highlight that within the backbone network of the landmark detection module, each instance of 3×3 convolution is substituted with the MSFM. The coordinates of each landmark are finally output.

4) MULTI-SCALE SEMANTIC FUSION MODULE

Due to the difference in original sample size and human anatomical structure, although the candidate region proposals obtained in RFAM are the same size, the local receptive field is still different. Candidate region proposals from different local receptive fields may have different levels of semantic information. In order to bridge the semantic gap between these candidate region proposals, we introduce MSFM, which uses feature maps from multiple scales and fuses them to produce semantically rich features. It can effectively extract spatial information from different scales, thus merging adjacent scales of context features more accurately and establishing long-range channel dependence. This approach ensures rich semantic information at all levels and can be efficiently implemented with a single input image scale.

To achieve this, we adopt a multi-branch approach to extract spatial information from the input feature map, with each branch having a channel dimension of C. We process tensors of multiple scales in parallel, allowing us to acquire richer positional information with varying spatial resolutions and depths. Specifically, the input of this part is the characteristic diagram after the convolution layer, each feature map with different scales, denoted as F_i, shares a common channel dimension $C' = C/S$, where $i = 0, 1, \cdots, S-1$. It is important

to note that C should be divisible by S. The generating function of the multi-scale feature map is expressed as:

$$F_i = Conv\left(k_i \times k_i, G_i = 2^{\frac{k_i-1}{2}}\right)(X)\, i = 0, 1, 2, \cdots, S-1 \tag{8}$$

The feature map with different scales, denoted as $F_i \in R^{C' \times H \times W}$, is obtained using the ith kernel size $k_i = 2 \times (i+1) + 1$. After processing tensors of multiple scales in parallel, we obtain the final combined feature map, which is as follows:

$$F = Cat\left([F_0, F_1, \cdots, F_{S-1}]\right) \tag{9}$$

The method for weighting the obtained multi-scale combined feature map is as follows, namely GAP + extraction

$$Z_c = \sigma(W_1\delta(W_0(\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_c(i,j)))) \tag{10}$$

Here, Z_c is the weight vector, and the symbol δ represents the rectified linear unit (ReLU) activation function. And $c = 0, 1, 2, \cdots, S-1$.The matrices $W_0 \in \mathbb{R}^{C \times \frac{C}{r}}$ and $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ represent the fully-connected (FC) layers. The symbol σ stands for the sigmoid activation function. H and W represent the height and width dimensions, respectively, and F_c represents the feature map of different channels. To achieve the interaction of attention information, the multi-scale channel attention vectors are computed and combined to obtain the overall multi-scale weight vector Z. Then the right of appeal is re-assigned by softmax as:

$$\bar{Z}_c = \frac{\exp(Z_c)}{\sum_{i=0}^{S-1}\exp(Z_c)} \tag{11}$$

Similarly, the new weight vector, denoted as $\bar{Z}$, is obtained through a series of connections of the multi-scale channel attention vectors. Then, we perform element-wise multiplication between the newly calibrated weight vector $\bar{Z}$ and the feature map F of the corresponding scale to obtain its feature map Y. The final output $\bar{Y}$ is obtained by concatenating the feature maps Y_i from each MSFM. Specifically, all the depth channels of the output feature maps are stacked together along the channel direction to obtain the final depth convolution result $\bar{Y}$. By combining information from multiple blocks, the MSFM can capture both fine-grained, local feature semantics from high-resolution feature maps and coarse, global feature semantics from low-resolution feature maps. This helps refine the landmark localization and produce more accurate results.

C. LOSS FUNCTION

The adaptive wing loss (AWL) [40] is used to train the framework. During the training period, AWL gradually reduces the influence of pelvic foreground pixels, that is, weakly related semantic information until the error becomes very small. So as to pay more attention to the strongly related semantic information around landmarks and make the model converge faster. The AWL is presented below:

$$\mathrm{L} = \begin{cases} \omega \ln\left(1 + |\frac{y-\hat{y}}{\epsilon}|^{\alpha-y}\right) & \text{if } |\left(y-\hat{y}\right)| < \theta \\ A\,|\,y-\hat{y}\,| - C & \text{otherwise} \end{cases} \tag{12}$$

Among them, y and $\hat{y}$ respectively ground truth and predicted heatmap. Here,

$$A = \omega(1/(1+(\frac{\theta}{\epsilon})^{(\alpha-y)}))(\alpha-y)((\frac{\theta}{\epsilon})^{(\alpha-y-1)})(\frac{1}{\epsilon}) \tag{13}$$

$$C = (\theta A - \omega \ln(1+(\theta/\epsilon)^{\alpha-y})) \tag{14}$$

Here, $\omega = 14$, $\theta = 0.5$, $\epsilon = 1$, and $\alpha = 2.1$ are positive numbers.

IV. EXPERIMENTS AND RESULTS

A. DATASET

To our knowledge, there is no public and available pelvic landmark dataset on X-rays. In order to conduct this study, we collected 430 pelvic X-ray images from the Shandong Provincial Hospital of Traditional Chinese Medicine. The collection consists of pelycograms obtained from 430 patients with resolutions ranging from 1670 × 2010 pixels to 3200 × 3200 pixels, with a spatial resolution of 0.139 mm/pixel in both directions. The patients included in this dataset span a wide age range, from 10 to 80 years old. The data set was manually marked with 17 landmarks on the X-ray image by two professional doctors with Make Sense software, and the average value of the two professional doctors was taken as the label. In our experiments, we divided the dataset into three sets, i.e. training set, validation set, and testing set in the ratio of 300, 30, and 100, respectively.

B. DATA AUGMENTATION

Unlike other areas of computer vision, the datasets for medical image analysis are often limited in size, as the process of obtaining ground-truth labels from clinical experts is resource-intensive and time-consuming. Consequently, training deep learning models on such a small amount of data can lead to overfitting and result in poor performance in clinical applications. To mitigate this issue, data augmentation has emerged as a powerful method [41]. Given that pelvic radiographs hold rich and complex morphometric information, it is important to use specialized data augmentation techniques, as demonstrated by Maini et al. [42], to ensure that the synthetic data remains clinically relevant and useful for training robust models. In our research, we focus on the pelvic X-ray anatomical structure and the time diversity. The following special image transformations are adopted, as elaborated below.

1) REFLECTION

Radiographs are subjected to a random flip, and subsequently, the corresponding landmarks are translated as:

$$\mathbf{L}_x^{(i)} = W - \mathbf{L}_x^{(i)}$$

$$\mathbf{L}_y^{(i)} = H - \mathbf{L}_y^{(i)} \tag{15}$$

2) UNSHARP MASKING

Considering the time diversity of samples, to enhance contrast and sharpness, we utilized a linear filter that selectively amplifies the high-frequency content of radiographs. The applied method can be represented by

$$I_{sharp} = I_{orig} + \alpha * \left(I_{orig} - I_{orig} * F\right) \tag{16}$$

where F is a linear filter and α is the sharpness amount.

3) SOLARIZATION

Considering the complexity of the anatomical structure of the sample, we utilized the solarization method to mitigate tone overexposure in radiographs and enhance the visibility of complex pelvic structures.

C. IMPLEMENTATION DETAILS

Training in this study was executed utilizing a TITAN RTX 24G GPU. The experimentation process was facilitated using the deep learning framework PyTorch. The images were resized to dimensions of 512 × 512 pixels within the backbone candidate region proposal extractor. The original dimensions were retained within the receptive field amplification module, while the candidate region proposals were resized to dimensions of 256 × 256 pixels within the landmark detection module. We set the data augmentation probability to 0.1. The backbone candidate region proposal extractor and landmark detection module are trained separately. For training these two modules, we both employed the Adam optimizer. Similarly, in a total of 100 cycles of training, the initial learning rate was set at 0.001, with subsequent reduction by a factor of 0.80 every 10 cycles. The loss functions are all AWL.

D. EVALUATION INDICES

In this paper, we evaluate the detection performance using these evaluation indices as recommended by the 2015 ISBI challenges on cephalometric landmark detection [43]. We evaluate landmark detection performance using the mean radial error (MRE). The smaller the MRE, the higher the detection accuracy. The MRE is calculated using

$$\text{MRE} = \frac{\sum_{i=1}^{k} \sqrt{\Delta x^2 + \Delta y^2}_i}{k} \tag{17}$$

Δx and Δy are the absolute differences between the estimated and ground-truth coordinates of the x- and y-axes. Similarly, D_i is defined as the difference between the real distance and the predicted distance and n is the number of images, the mean distance error (MDE) for clinical measurements is defined as:

$$\text{MDE} = \frac{\sum_{i=1}^{n} \mathrm{D}_i}{n} \tag{18}$$

To account for the differences between predicted results and the ground truth, we define a certain range within which a prediction is considered correct. Specifically, we evaluate the range of z mm (where z = 4, 4.5, 5, and 6) in our experiment. For instance, if the radial error is 3.5 mm, we consider it a success within the 4 mm range. The successful detection rate (SDR) is defined as the percentage of landmarks detected successfully within the specified range. The larger the SDR, the better the performance of the proof model. Equation 18 provides a formal definition of SDR:

$$\text{SDR} = \frac{\mathrm{N_a}}{\mathrm{N}} \times 100\% \tag{19}$$

where N_a indicates the number of accurate detections, and N indicates the total number of detections.

In addition, we further evaluate the effect of the model landmark predictions on pelvic angle clinical measurement. Symmetric mean absolute percentage (SMAPE) evaluates overall performance on angle measurement. The SMAPE metric is defined as:

$$SMAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{SUM\,|X_i - Y_i\,|}{SUM\,(X_i + Y_i)} \times 100\% \tag{20}$$

where X_i means the predicted angles and Y_i means the ground truth. A smaller value of SMAPE means more accurate angle predictions.

E. RESULTS

We evaluated the performance of our proposed pelvic landmark detection framework on our test datasets, which consist of 100 pelvic images. Table 1 summarizes the results in terms of MRE with standard deviation (STD) for all 17 landmarks on the test set, as well as the SDR within four clinical ranges (i.e. 4.0mm, 4.5mm, 5.0mm, and 6.0mm) for each landmark. The * denotes outcomes achieved by retraining our dataset through the process of replicating the original paper's methodology. The remaining results are drawn directly from the original paper, offering a comprehensive foundation for comparative analysis. To maintain consistency, all measurements have been converted to millimeters (mm), with a spatial resolution of 0.139 mm/pixel employed as the default for experiments lacking explicit spatial resolution specifications. The average MRE with STD across all landmarks for 100 test images is evaluated to be 3.724 ± 4.247mm, which falls within the clinically accepted precision range of 4.0mm. These results indicate that our framework is capable of locating cephalometric landmarks accurately and consistently, as evidenced by the smaller MRE and STD values. In the test set, the MRE varies from 1.737mm (L15) to 7.140mm (L2), while the SDR ranges from 47% to 96% within a 4.0mm neighborhood. On average, all landmarks exhibit SDRs of 74.176%, 78.117%, 80.588%, and 84.706% for neighborhoods of 4.0mm, 4.5mm, 5.0mm, and 6.0mm, respectively. Moreover, our proposed framework exhibits robust and effective performance in detecting most landmarks. Notably, landmarks L3, L4, L5, L10, L12, L13, L14, and L15 yield high SDR values, further attesting to the reliability of our approach.

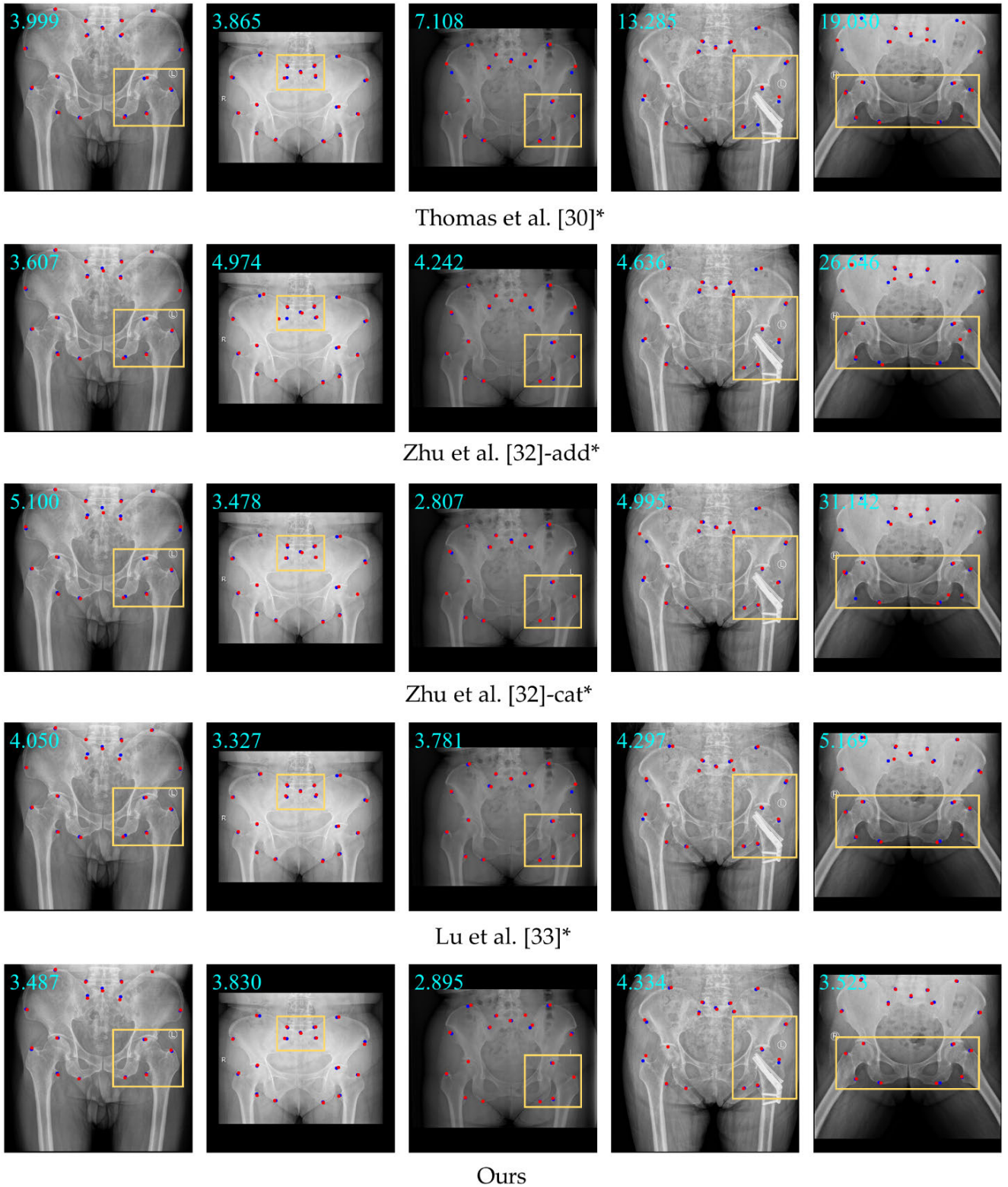


FIGURE 4. Examples of predictions from different methods. Automatically located (red) and manually annotated pelvic landmarks (blue) in the five subjects. The yellow boxes denote significant disparities in the experimental results. Furthermore, the MRE for each sample is indicated in the upper left corner of its respective image.

At the same time, as shown in Table 2, compared with the most advanced methods, our proposed framework shows the most advanced performance. Our framework achieves the best MRE of 3.724mm and the best STD of 4.247mm among

TABLE 1. The final detection result of each landmark. The results are presented in terms of MRE, STD, and SDR within 4.0, 4.5, 5.0, and 6.0 mm neighborhoods for all 17 landmarks. MRE is reported in millimeters (mm), and SDR values are reported as a percentage (%).

Landmarks	SDR (%)				MRE&STD
	4mm	4.5mm	5mm	6mm	
L1	55	62	64	72	5.169 ± 4.582
L2	47	52	54	60	7.140 ± 6.094
L3	91	92	92	92	2.857 ± 5.874
L4	91	92	92	92	2.871 ± 5.825
L5	83	87	90	91	3.023 ± 4.763
L6	68	72	77	85	4.039 ± 4.026
L7	65	72	80	87	4.385 ± 6.065
L8	67	71	73	77	4.477 ± 4.740
L9	55	60	65	70	5.512 ± 5.450
L10	81	85	85	90	2.689 ± 2.293
L11	53	61	61	73	4.180 ± 2.955
L12	88	93	93	95	2.065 ± 2.396
L13	86	87	87	93	2.982 ± 3.885
L14	96	97	97	98	1.828 ± 1.677
L15	96	97	97	98	1.737 ± 1.286
L16	70	75	75	86	3.914 ± 4.923
L17	69	73	74	80	4.432 ± 5.247
Mean	**74.176**	**78.117**	**80.588**	**84.706**	**3.724 ± 4.247**

TABLE 2. Comparison of our proposed framework with state-of-the-art methods in terms of MRE, STD, MDE, and SMAPE. On our dataset, the best results are highlighted in bold, and the second runner-up is underlined. The table clearly shows that our proposed framework outperformed the state-of-the-art in MRE, STD, MDE, and SMAPE on the test set.

Methods	MRE (mm)	STD (mm)	MDE (mm)	SMAPE (%)
[30]	2.450	0.640	/	/
[30]*	5.328	<u>5.507</u>	<u>5.834</u>	30.361
[5]	5.600	4.500	/	/
[32]-add	7.226	21.140	/	/
[32]-add*	6.183	19.711	9.675	31.273
[32]-cat	4.669	12.552	/	/
[32]-cat*	6.419	20.499	9.869	32.240
[33]*	<u>4.159</u>	5.015	6.067	<u>28.988</u>
Ours	**3.724**	**4.335**	**5.262**	**27.846**

all methods in our test set. In addition, our framework also achieved the best MDE of 5.262mm and the best SMAPE of 27.864%. When scrutinizing the experimental outcomes of these state-of-the-art methods across both their proprietary datasets and our dataset, it becomes evident that the MRE and STD obtained by these methods on our dataset have notably increased.

It becomes evident that our method not only demonstrates commendable accuracy but also exhibits robustness on the intricate dataset presented in this paper. Despite the fact that our pelvis sample size surpasses that of Zhu et al. [32], the MRE has experienced an increase of 1.75mm, and the STD has similarly risen by 7.947mm. This once again underscores the challenge of effectively learning complex pelvic X-ray data in real clinical scenarios. Additionally, it's noteworthy that the existing methods primarily focus on grasping the global semantic context of pelvic X-ray images, whereas our approach specifically focuses on acquiring strongly related local semantic information surrounding landmarks. As hypothesized, our supposition holds true: that when it comes to detecting landmarks within intricate pelvic X-ray samples, the presence of strongly related semantic information within an expansive receptive field enhances the model's convergence performance.

To facilitate a more intuitive comparison of the detection outcomes across various methods, we present visualizations of the detection results on several samples from this dataset. Illustrated in Figure 4, each column represents an identical sample, and each row corresponds to the outcome produced by a specific method. Obviously, our method has achieved satisfactory prediction results in both irregular and regular pelvises. Especially in the areas marked in the yellow box, our method is obviously improved compared with other advanced methods. Most of the landmarks in these locations are typical representatives of time diversity and pathological abnormalities. Obviously, our method has better robustness when detecting these landmarks. Fig. 5 shows the box-and-whisker plots of point to point error for each method. The range of distance error for each point in our method is more concentrated. Although our method gets small errors on most points, it performs poorly on landmark 2. We can also see that the detection ability of other methods has been greatly affected by the variety of data.

In conclusion, our proposed framework demonstrates the potential to be employed in real-world applications for accurate pelvic landmark detection. Compared with the most advanced methods, the overall MRE and STD obtained by our method in landmark detection are satisfactory, especially on samples with time diversity and pathological abnormalities. In addition, the evaluation results of clinical measurements in pelvic abnormality examinations verified the effectiveness of our method once again.

V. DISCUSSION

The accurate localization of anatomical landmarks in pelvic images is essential for the examination, treatment, and prognosis of pelvic abnormalities. However, the X-ray of the pelvis in clinical practice is multi-scale, with time diversity and pathological deformity. The intricate relationship between image features and landmark positions makes it difficult to achieve precise and reliable results, which makes it a challenging task to detect pelvic landmarks.

We conducted extensive experiments to evaluate the performance of our proposed framework against state-of-the-art methods on our irregular dataset. Our results demonstrate significant improvements over existing approaches, showcasing the effectiveness of our solution. Specifically, our framework achieves an MRE of 3.714 ± 4.247mm and the highest SDRs of 74.176%, 78.117%, 80.588%, and 84.706% in clinical precision ranges of 4.0mm, 4.5mm, 5.0mm, and 6.0mm, respectively. At the same time, in the evaluation of clinical measurement values, the detection results obtained by our framework can achieve an MDE of 5.262mm and a SMAPE of 27.846%.

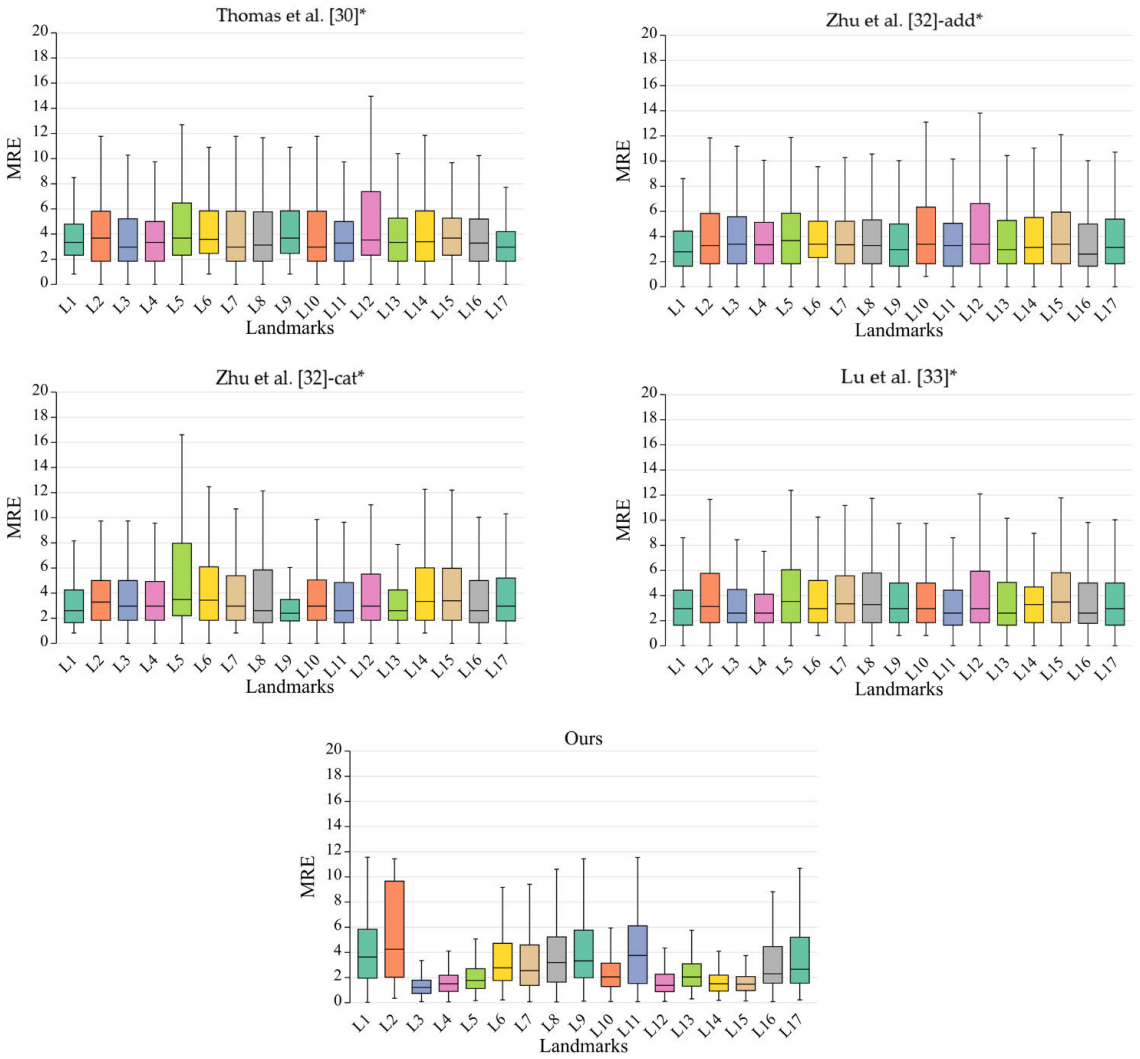


FIGURE 5. The Box-and-whisker plots of Euclidean distance distribution between predicted landmarks and ground truths.

A. ABLATION STUDIES

In this section, we conduct a comprehensive analysis of the proposed framework, exploring its performance and effectiveness from various perspectives. We first aim to evaluate the impact of different backbone networks on the framework's detection capabilities. Specifically, we validate the performance of our framework using a range of backbone architectures, including SHG [44], HRNet [45], CE-net [46], Resnet50 [47], and U-net. Table 3 provides a comprehensive comparison of different backbone networks in the backbone candidate region proposal extractor and landmark detection module. The results highlight that U-net outperforms other backbone networks in the backbone candidate region proposal extractor, exhibiting the lowest MRE of 5.963 ± 8.311mm. That is, the actual average error value is 42.899px, and the actual theoretical maximum error value is 102.69px, which meets the requirement of less than s/2 mentioned in Section III-B1. Similarly, within the landmark detection module, Resnet50 emerges as the optimal choice. Here, the experiments related to the landmark detection module are all carried out on the basis that the backbone of the backbone candidate region proposal extractor is U-net. However, in the absence of RFAM and MSFM, the performance of diverse backbone architectures even falls short of that of the backbone candidate region proposal extractor. Subsequently, we added RFAM and MSFM to Resnet50 sequentially, and the MRE decreased to 5.210 ± 6.307mm and 3.724 ± 4.247mm, respectively. Clearly, when confronted with the intricate pelvic dataset in this study, both the conventional end-to-end and two-stage frameworks struggle to attain high-precision detection. The conducted ablation experiment corroborates the effectiveness of our RFAM and MSFM. This affirms the validity of our concept concerning the expansion of the receptive field and the integration of multi-scale semantic fusion.

Key point detection differs significantly from target detection. Common techniques used in target detection, such as R-CNN [48], [49] and YOLO [50], [51], follow a top-down approach, where the category is initially determined before detection. Some studies have attempted to apply YOLOv3

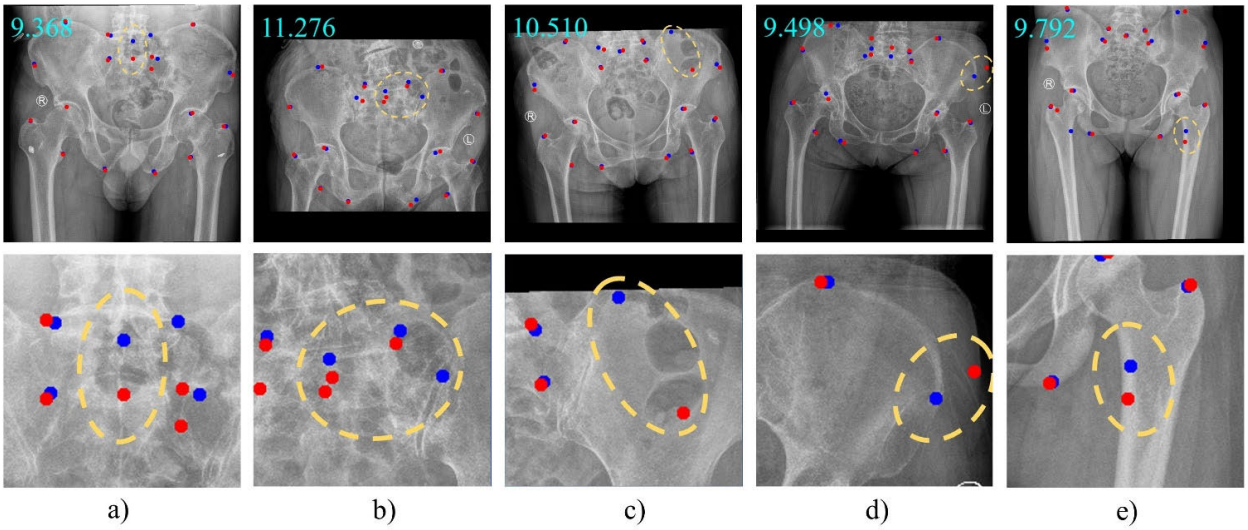


FIGURE 6. The landmark with significant detection failure is marked with a yellow dotted box, and below it is a local enlarged view of the surrounding area.

TABLE 3. Performance comparison of different backbone networks in the backbone candidate region proposal extractor and landmark detection module. Notably, our proposed RFAM and MSFM both performed well, reducing MRE by 6.492mm and 1.486mm, respectively.

Backbone candidate region proposals extractor		
Methods	MRE (mm)	STD (mm)
SHG (4 stages)	9.247	9.729
HRNet	7.262	8.879
Resnet50	6.464	8.706
CE-net	6.528	16.006
U-net	**5.963**	**8.311**
Landmark detection module		
Methods	MRE (mm)	STD (mm)
U-net	25.478	11.994
CE-net	21.283	10.489
Resnet50	11.702	6.191
Resnet50+RFAM	5.210	6.307
Resnet50+RFAM+MSFM	**3.724**	**4.247**

[52] to key point detection, but the results have been only moderately successful. In the context of key point detection, the top-down method may not always be suitable for various tasks. Instead, a bottom-up approach is often more appropriate, involving the direct identification of key points [45]. That is, the landmark detection stage after we get the candidate area.

B. FAILURE ANALYSIS

In our proposed framework, challenges beyond the multi-scale characteristics, time diversity, and pathological irregularities of pelvic X-rays include occlusions and misalignments of edge landmarks. These factors significantly impede the performance of our framework. To delve into this matter, we conducted an analysis of instances where landmarks were predicted with substantial errors.

In addition, the labeling of landmarks in data sets is variable and subjective. In fact, the difference between the two experts' labels on landmarks reached 2.14 ± 1.57 mm, and our first evaluation range was 4 mm. This leads to a higher degree of uncertainty in the real landmarks on the ground, and ultimately affects the performance of our proposed framework.

Figure 6 illustrates five examples featuring predicted landmarks with a high MRE, exceeding the clinically acceptable range. As shown in Figure 6.a, the adjacent bones below L5 have extremely similar anatomical structures, which disturbs the convergence of the model and leads to a suboptimal result. In Figure 6.b, the image itself is too blurred, which leads to a large detection error. Figure 6.c shows a few incomplete shots, while Figure 6.d demonstrates that the bone edge point shares a shape resemblance with the soft tissue edge point, further contributing to inaccuracies. Lastly, Figure 6.e showcases the unclear actual position of L15. All these scenarios contribute to no-table final detection errors. To address this issue, we may explore the integration of additional pelvic features or image augmentation techniques to improve the robustness of our framework to handle these complicated situations.

C. CURRENT CHALLENGES AND FUTURE DIRECTIONS

The results presented in sections IV and V clearly demonstrate the significant contribution made by our proposed framework towards automated pelvic analysis, achieving state-of-the-art level results on our dataset. However, despite this achievement, there remain several challenges that require further attention.

One of the primary challenges in developing an accurate pelvic landmark detection model is the limited size of the available training dataset. In our dataset, the training data consists of only 300 pelvic images ranging in age from 10 to 80 years old. This limited sample size and diverse patient pool can make it difficult for an AI algorithm to effectively generalize and may lead to overfitting. In addition, as

mentioned earlier, the labeling of landmarks is subjective, and the labeling gap between the two experts is obvious. Therefore, there is a need for new state-of-the-art datasets that can help overcome these challenges.

Similarly, the trend toward constructing single-training models for multimodal data has gained momentum. Our method can't solve double branch attention driven multi-scale learning for MRI for the time being, which is not only a challenge but also the direction of our follow-up work. Nonetheless, the preservation of high performance for such extensive models when confronted with medical images, particularly within the intricate dataset outlined in this paper, remains a formidable obstacle. Addressing this challenge necessitates researchers to delve further into the latent information existing both between and within modes, thereby unlocking the full potential embedded within these complexities.

VI. CONCLUSION

This research addresses the crucial need for accurate pelvic landmark detection in abnormal pelvic examinations, offering a promising solution to automate the tracing process and enhance clinical efficiency. By introducing a two-stage regression framework with RFAM and MSFM, we have significantly overcome the limitations of existing approaches, which can't fully learn the strong related semantic information of the large receptive field at high resolution. By combining multi-scale and semantically rich features, our framework makes full use of the strongly relevant semantic information around landmarks and provides a comprehensive understanding of pelvic abnormalities. The results obtained from our framework showcase its potential to revolutionize pelvic analysis by reducing subjectivity and the time required for manual landmark identification. We anticipate that our proposed framework will assist in improving patient outcomes, advancing treatment strategies, and facilitating a comprehensive assessment of pelvic abnormalities.

Of course, our method can be used to realize clinical computer-aided diagnosis (CAD) and support the activities of clinicians, but it needs to strengthen the constraints on security and service availability [53]. In the clinical environment, the continuity of service is a mandatory requirement, and the use of service-oriented networks (SONs) can improve the continuity. Benefiting from [54], we can realize son-based services in clinical scenarios.

Our future work will not only improve the efficiency of this method but also extend it to a multimodal general large model that only needs one training. In addition, we will continue to collect data sets, and the first publicly available pelvic X-ray landmark detection data set will be established later.

ACKNOWLEDGMENT

The pelvis X-ray test dataset was provided by Fangwei Xu from the Shandong Hospital of Traditional Chinese Medicine, as well as her team members. The authors would like to thank Haifeng Wei from the Shandong Hospital of Traditional Chinese Medicine for providing the pelvis X-ray training dataset.

REFERENCES

[1] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021, doi: 10.1109/JPROC.2021.3054390.

[2] D. Yang, S. Zhang, Z. Yan, C. Tan, K. Li, and D. Metaxas, "Automated anatomical landmark detection ondistal femur surface using convolutional neural network," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 17–21.

[3] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H. K. Huang, "Bone age assessment of children using a digital hand atlas," *Computerized Med. Imag. Graph.*, vol. 31, nos. 4–5, pp. 322–331, Jun. 2007.

[4] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, "Robust landmark detection in volumetric data with efficient 3D deep learning," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing* (Advances in Computer Vision and Pattern Recognition), L. Lu, Y. Zheng, G. Carneiro, L. Yang, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-42999-1_4.

[5] B. Bier, M. Unberath, J. N. Zaech, J. Fotouhi, M. Armand, G. Osgood, N. Navab, and A. Maier, "X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018* (Lecture Notes in Computer Science), A. Frangi, J. Schnabel, C. Davatzikos, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 55–63, doi: 10.1007/978-3-030-00937-3_7.

[6] T. Lange, N. Papenberg, S. Heldmann, J. Modersitzki, B. Fischer, H. Lamecker, and P. M. Schlag, "3D ultrasound-CT registration of the liver using combined landmark-intensity information," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 4, no. 1, pp. 79–88, Jan. 2009.

[7] Q. Yao, L. Xiao, P. Liu, and S. K. Zhou, "Label-free segmentation of COVID-19 lesions in lung CT," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2808–2819, Oct. 2021.

[8] N. Alexander, A. Rastelli, T. Webb, and D. Rajendran, "The validity of lumbo-pelvic landmark palpation by manual practitioners: A systematic review," *Int. J. Osteopathic Med.*, vol. 39, pp. 10–20, Mar. 2021, doi: 10.1016/j.ijosm.2020.10.008.

[9] M. Piron, L. Pop, V. Radoi, N. Bacalbasa, I. Balescu, and I. D. Suciu, "The Yabuki space: Landmark in pelvic anatomy and surgery," *Romanian Med. J.*, vol. 69, no. S3, pp. 36–37, Jun. 2022, doi: 10.37897/rmj.2022.s3.12.

[10] V. Grau, M. Alcañiz, M. C. Juan, C. Monserrat, and C. Knoll, "Automatic localization of cephalometric landmarks," *J. Biomed. Informat.*, vol. 34, no. 3, pp. 146–156, Jun. 2001.

[11] I. El-Feghi, M. A. Sid-Ahmed, and M. Ahmadi, "Automatic localization of craniofacial landmarks for assisted cephalometry," *Pattern Recognit.*, vol. 37, no. 3, pp. 609–621, Mar. 2004.

[12] J. Keustermans, W. Mollemans, D. Vandermeulen, and P. Suetens, "Automated cephalometric landmark identification using shape and local appearance models," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2464–2467.

[13] C. Lindner, C.-W. Wang, C.-T. Huang, C.-H. Li, S.-W. Chang, and T. F. Cootes, "Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms," *Sci. Rep.*, vol. 6, no. 1, p. 33581, Sep. 2016.

[14] S. Ö. Arik, B. Ibragimov, and L. Xing, "Fully automated quantitative cephalometry using convolutional neural networks," *J. Med. Imag.*, vol. 4, no. 1, Jan. 2017, Art. no. 014501.

[15] S. Park, "Cephalometric landmarks detection using fully convolutional networks," M.S. thesis, Colleage Natural Sci., Seoul Nat. Univ, Seoul, South Korea, 2017.

[16] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Image Anal.*, vol. 54, pp. 207–219, Jan. 2019.

[17] R. Chen, Y. Ma, N. Chen, D. Lee, and W. Wang, "Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 873–881.

[18] J. Qian, W. Luo, M. Cheng, Y. Tao, J. Lin, and H. Lin, "CephaNN: A multi-head attention network for cephalometric landmark detection," *IEEE Access*, vol. 8, pp. 112633–112641, 2020, doi: 10.1109/ACCESS.2020.3002939.
[19] H. Wu, C. Bailey, P. Rasoulinejad, and S. Li, "Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science), vol. 10433, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. Collins, S. Duchesne, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-66182-7_15.
[20] M. Šavc, G. Sedej, and B. Potočnik, "Cephalometric landmark detection in lateral skull X-ray images by using improved SpatialConfiguration-Net," *Appl. Sci.*, vol. 12, no. 9, p. 4644, May 2022, doi: 10.3390/app12094644.
[21] Z. Zhong, J. Li, Z. Zhang, Z. Jiao, and X. Gao, "An attention-guided deep regression model for landmark detection in cephalograms," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 540–548.
[22] H. Kim, E. Shim, J. Park, Y.-J. Kim, U. Lee, and Y. Kim, "Web-based fully automated cephalometric analysis by deep learning," *Comput. Methods Programs Biomed.*, vol. 194, Oct. 2020, Art. no. 105513, doi: 10.1016/j.cmpb.2020.105513.
[23] Y. Song, X. Qiao, Y. Iwamoto, and Y.-W. Chen, "Automatic cephalometric landmark detection on X-ray images using a deep-learning method," *Appl. Sci.*, vol. 10, no. 7, p. 2547, Apr. 2020, doi: 10.3390/app10072547.
[24] F. Schwendicke, A. Chaurasia, L. Arsiwala, J.-H. Lee, K. Elhennawy, P.-G. Jost-Brinkmann, F. Demarco, and J. Krois, "Deep learning for cephalometric landmark detection: Systematic review and meta-analysis," *Clin. Oral Investigations*, vol. 25, no. 7, pp. 4299–4309, Jul. 2021, doi: 10.1007/s00784-021-03990-w.
[25] J. Li, Y. Wang, and G. Li, "Research and challenges of medical image landmark detection based on deep learning," *Acta Electonica Sinica*, vol. 50, no. 1, p. 226, 2022.
[26] J. Yuan, Z. Deng, S. Wang, and Z. Luo, "Multi receptive field network for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1883–1892.
[27] M. Aggarwal, V. Khullar, N. Goyal, A. Alammari, M. A. Albahar, and A. Singh, "Lightweight federated learning for rice leaf disease classification using non independent and identically distributed images," *Sustainability*, vol. 15, no. 16, p. 12149, Aug. 2023, doi: 10.3390/su151612149.
[28] Y. Pei, L. Mu, C. Xu, Q. Li, G. Sen, B. Sun, X. Li, and X. Li, "Learning-based landmark detection in pelvis X-rays with attention mechanism: Data from the osteoarthritis initiative," *Biomed. Phys. Eng. Exp.*, vol. 9, no. 2, Mar. 2023, Art. no. 025001, doi: 10.1088/2057-1976/ac8ffa.
[29] F.-Y. Liu, C.-C. Chen, C.-T. Cheng, C.-T. Wu, C.-P. Hsu, C.-Y. Fu, S.-C. Chen, C.-H. Liao, and M. S. Lee, "Automatic hip detection in anteroposterior pelvic radiographs—A labelless practical framework," *J. Personalized Med.*, vol. 11, no. 6, p. 522, Jun. 2021.
[30] T. Statchen, G. Choi, C. Gao, W. Xu, C. Rajapakse, L. Hao, D. Ranaweera, C. J. Lee, E. Tu, and A. Basu, "Automated femoral landmark detection from pelvic X-rays," *J. Clin. Densitometry*, vol. 25, no. 2, p. 282, Apr. 2022.
[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015* (Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. Wells, A. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[32] H. Zhu, Q. Yao, L. Xiao, and S. K. Zhou, "Learning to localize cross-anatomy landmarks in X-ray images with a universal model," *BME Frontiers*, vol. 2022, Jan. 2022, Art. no. 9765095, doi: 10.34133/2022/9765095.
[33] C. Lu, W. Chen, X. Qiao, and Q. Zeng, "Prior active shape model for detecting pelvic landmarks," *Innovation in Medicine and Healthcare* (Smart Innovation, Systems and Technologies), vol. 357. Singapore: Springer, 2023, doi: 10.1007/978-981-99-3311-2_26.
[34] J. Qian, M. Cheng, Y. Tao, J. Lin, and H. Lin, "CephaNet: An improved faster R-CNN for cephalometric landmark detection," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 868–871.
[35] J. H. Lee, H. J. Yu, M. Kim, J. W. Kim, and J. Choi, "Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks," *BMC Oral Health*, vol. 20, no. 1, pp. 1–10, Oct. 2020.
[36] M. Lee, M. Chung, and Y.-G. Shin, "Cephalometric landmark detection via global and local encoders and patch-wise attentions," *Neurocomputing*, vol. 470, pp. 182–189, Jan. 2022, doi: 10.1016/j.neucom.2021.11.003.
[37] M. Zeng, Z. Yan, S. Liu, Y. Zhou, and L. Qiu, "Cascaded convolutional networks for automatic cephalometric landmark detection," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101904.
[38] H. J. Kwon, H. I. Koo, J. Park, and N. I. Cho, "Multistage probabilistic approach for the localization of cephalometric landmarks," *IEEE Access*, vol. 9, pp. 21306–21314, 2021.
[39] Y. Ao and H. Wu, "Feature aggregation and refinement network for 2D anatomical landmark detection," *J. Digit. Imag.*, vol. 36, no. 2, pp. 547–561, Nov. 2022, doi: 10.1007/s10278-022-00718-4.
[40] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6970–6980.
[41] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[42] R. Maini and H. Aggarwal, "A comprehensive review of image enhancement techniques," 2010, *arXiv:1003.4053*.
[43] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, P. Fischer, T. F. Cootes, and C. Lindner, "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.*, vol. 31, pp. 63–76, Jul. 2016.
[44] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 483–499, doi: 10.1109/CVPR46437.2021.01584.
[45] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696, doi: 10.1109/CVPR.2019.00584.
[46] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019, doi: 10.1109/TMI.2019.2903562.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[48] J. Xu, H. Ren, S. Cai, and X. Zhang, "An improved faster R-CNN algorithm for assisted detection of lung nodules," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106470, doi: 10.1016/j.compbiomed.2022.106470.
[49] M. E. Sahin, H. Ulutas, E. Yuce, and M. F. Erkoc, "Detection and classification of COVID-19 by using faster R-CNN and mask R-CNN on CT images," *Neural Comput. Appl.*, vol. 35, no. 18, pp. 13597–13611, Jun. 2023.
[50] F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio, and S. Vitabile, "A YOLO-based model for breast cancer detection in mammograms," *Cognit. Comput.*, vol. 2023, pp. 1–14, Aug. 2023, doi: 10.1007/s12559-023-10189-6.
[51] N. Aishwarya, K. Manoj Prabhakaran, F. T. Debebe, M. S. S. A. Reddy, and P. Pranavee, "Skin cancer diagnosis with YOLO deep neural network," *Proc. Comput. Sci.*, vol. 220, pp. 651–658, Jan. 2023.
[52] C.-H. King, Y.-L. Wang, W.-Y. Lin, and C.-L. Tsai, "Automatic cephalometric landmark detection on X-ray images using object detection," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Kolkata, India, Mar. 2022, pp. 1–4, doi: 10.1109/ISBI52829.2022.9761506.
[53] J. Yanase and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Syst. Appl.*, vol. 138, Dec. 2019, Art. no. 112821, doi: 10.1016/j.eswa.2019.112821.
[54] V. Conti, C. Militello, L. Rundo, and S. Vitabile, "A novel bio-inspired approach for high-performance management in service-oriented networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 4, pp. 1709–1722, Oct. 2021, doi: 10.1109/TETC.2020.3018312.

CHENYANG LU was born in Henan, China, in 2000. He received the bachelor's degree from Henan Polytechnic University, Jiaozuo, China, in 2022. He is currently pursuing the master's degree in electronic information with the School of Control Science and Engineering, Shandong University. His research interests include medical image processing and deep learning.

JIAGENG ZHAO was born in 1999. He is currently pursuing the master's degree in acupuncture, moxibustion, and massage with the Shandong University of Traditional Chinese Medicine. He has published two articles in national journals. He is good at using traditional Chinese medicine to treat traumatology and internal medicine diseases.

WEI CHEN (Member, IEEE) received the B.S. degree in electrical engineering and its automation and the M.S. degree in control science and engineering from Qufu Normal University, China, in 2012 and 2015, respectively, and the Ph.D. degree in biomedical engineering from Shandong University, China, in 2020. From 2020 to 2022, he was a Postdoctoral Researcher with the School and Hospital of Stomatology, Cheeloo College of Medicine, Shandong University. Since 2023, he has been an Associate Professor of medical imaging with the School of Radiology, Shandong First Medical University and Shandong Academy of Medical Sciences. His current research interests include deep learning and medical image analysis.

XU QIAO (Member, IEEE) received the B.S. degree in mathematical statistics and the M.S. degree from Shandong University, China, in 2004 and 2007, respectively, and the Ph.D. degree from Ritsumeikan University, Japan, in 2010. From 2010 to 2012, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS). Since 2018, he has been an Associate Professor of biomedical engineering with the School of Control Science and Engineering, Shandong University. His research interests include imaging diagnosis and medical image analysis.

QINGYUN ZENG was born in 1974. She received the master's and Ph.D. degrees in medicine, in 2003 and 2006, respectively. She has rich experience in orthopedics and internal medicine massage treatment. She is currently the Chief Physician of the Department of Tuina, Affiliated Hospital of Shandong University of Traditional Chinese Medicine. She has published multiple academic articles in national and provincial core journals.