

# A Deep-Learning-Based Lumbosacral Localization and Landmark Detection Network for Automatic Lumbar Stability and Spondylolisthesis Grading Assessment

Tingting Hu<sup>1</sup>, Rong Zhang<sup>\*1</sup>, Baolin Xu<sup>2</sup>, Dongdong Xia<sup>2</sup>, Qiang Li<sup>3</sup>, Lijun Guo<sup>1</sup>

<sup>1</sup> *The Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China*

<sup>2</sup> *The Department of Orthopedics, the First Affiliated Hospital of Ningbo University, Ningbo, China*

<sup>3</sup> *The Department of Radiology, the Affiliated People's Hospital of Ningbo University, Ningbo, China*

\* Corresponding author, e-mail: zhangrong@nbu.edu.cn

**Abstract**—The accurate detection of vertebral landmarks is crucial for clinical diagnosis and research on the lumbar stability and spondylolisthesis grading. However, the small size of vertebral landmarks in X-ray images and the morphological similarity among vertebrae complicate this detection task. Recent advances in deep learning have enhanced the spinal landmark detection. To further improve the assessment methods for the lumbar stability and spondylolisthesis grading, we propose the LSLLD-Net, a novel network based on the lumbosacral localization and landmark detection. In clinical practices, the X-ray images used for evaluating the lumbar spine stability can often include the extraneous information from non-lumbosacral areas, such as the thoracic spine. The proposed LSLLD-Net first extracts the lumbosacral region from the complex X-ray images and then performs the precise landmark detection within the identified area. The landmark detection stage integrates the HRNet and UNet architectures, effectively capturing the long-range contextual information, the overall structural layout, and the fine local details in lumbar X-ray images to optimize landmark detection. Additionally, we propose a Multi-Scale Attention Module that enhances the relevant features and suppresses the irrelevant ones through channel and spatial attention, thereby achieving precise landmark detection and improving the robustness of the network. The evaluations on the private and public BUU-LSPINE datasets indicate that the LSLLD-Net surpasses other state-of-the-art methods in landmark detection, enhancing the accuracy and efficiency of Sagittal Displacement and Intervertebral Space Angle measurements. This performance excels in assessing the lumbar stability and spondylolisthesis grading, offering the significant support to clinicians for early quantitative diagnosis and evaluation.

**Index Terms**—lumbar stability, spondylolisthesis grading, LSLLD-Net, lumbosacral localization, landmark detection

## I. INTRODUCTION

THE assessment of lumbar stability and spondylolisthesis grading is essential for diagnosing lumbar diseases and making surgical decisions in clinical practice [1]. Clinically, the relative Sagittal Displacement (SD) or Intervertebral Space Angle (ISA) changes in the vertebrae observed in the hyperflexion and hyperextension X-ray images are commonly used as the diagnostic criteria for assessing the lumbar stability, while the percentage of SD in the lateral X-ray images is

utilized to grade the lumbar spondylolisthesis. To evaluate the lumbar stability, this study uses the threshold of the vertebral SD exceeding 3 mm or the ISA change exceeding 12° as the indicators of lumbar instability [2]. The calculation of the SD (Fig. 1 (a)) involves drawing a line parallel to the line connecting the posterior upper and lower corners of the lower vertebra from the posterior lower corner of the upper vertebra and measuring the horizontal distance between both lines. The calculation of the ISA (Fig. 1 (b)) is defined as the angle between the line connecting the two corners of the lower endplate of the upper vertebra and the line connecting the two corners of the upper endplate of the lower vertebra. To grade the lumbar spondylolisthesis, we employ the Meyerding grading method commonly used in clinical practices [3], from Fig. 1 (c), the upper endplate of the lower vertebra is divided into four equal parts, and the grade of spondylolisthesis is determined by the percentage of SD of the upper vertebra. The displacement of less than 25% is classified as Grade I, 25%–50% as Grade II, 50%–75% as Grade III, and 75%–100% as Grade IV. These parameters are measured using landmarks in the lumbar X-ray images. Currently, the measurement of vertebral SD and ISA in clinical practices is primarily performed manually by physicians or with PACS system measurement tools, which are time-consuming, labor-intensive, and subject to the physician's influence [4], [5]. Therefore, the accurate landmark detection is a crucial step for the computer-aided diagnosis of lumbar diseases. The development of a highly accurate automated landmark detection method is essential for effectively diagnosing lumbar stability and grading spondylolisthesis.

In recent years, various methods for the automated landmark detection in the spine have been explored, with the deep learning-based methods demonstrating the significant potential [6]. These methods are primarily categorized into two types: two-stage and single-stage detection methods. Two-stage detection methods typically involve first detecting each vertebra and then performing landmark detection within those areas. For example, Nguyen et al. [7] propose a convolutional

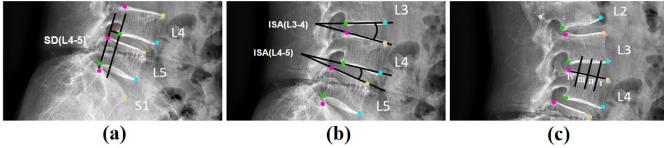


Fig. 1. Definitions of clinical parameters involved in the diagnosis of lumbar stability and spondylolisthesis grading: (a) definition of vertebral Sagittal Displacement; (b) definition of Intervertebral Space Angle; (c) definition of Meyerding's grading method.

neural network based on the VGG-net architecture to identify vertebral landmarks, followed by a secondary neural network to refine these landmarks for greater accuracy. Similarly, Zhang et al. [8] introduce a multitask learning framework that integrates the vertebral detection and landmark detection networks to achieve the high-precision results. The final accuracy of the landmark detection in two-stage methods relies on the precision of vertebral detection in the initial stage. In contrast, the single-stage detection methods generally include the coordinate regression-based, the heatmap regression-based, and the center-point with offset-based approaches. Liu et al. [9] employ the structural support vector regression networks to directly predict the spinal landmarks by utilizing the intrinsic geometric relationships among landmarks. Li et al. [10] introduce the Feature Decoupling and Gating Refinement Network, which applies a heatmap regression method to predict all landmarks. Yi J et al. [11] develop a U-net-based full spine landmark detection network that simultaneously predicts the center-point heatmaps, the center-point offsets, and the corner-point offsets to achieve the final landmark detection results.

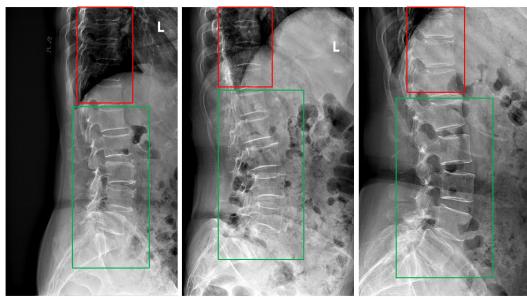


Fig. 2. Three examples of the lateral clinical X-ray images. Green boxes indicate lumbosacral areas, and red boxes indicate thoracic areas.

In the previous studies on vertebral landmark detection, the X-ray images typically covered the entire spine, where the number of vertebrae was fixed and the vertebral information was essentially what was required for tasks, making the processing relatively simpler. However, in the task of assessing lumbar stability and spondylolisthesis grading, the clinical X-ray images include the lumbosacral area (five lumbar vertebrae and one sacral vertebrae) within the green box and a variable number of thoracic vertebrae within the red box (Fig. 2). For the accurate assessment, only the vertebral information within the green box is needed. The extraneous information in the red box can interfere with the landmark detection task,

leading to the misalignment and the increased assessment difficulty. Furthermore, the full spine X-ray images provide comprehensive information about the entire spine, allowing the network to utilize rich contextual information for better vertebral location and landmark feature learning. However, the lumbosacral area only contains partial information about the spine. Consequently, with less contextual information available compared to the full-spine X-ray images, achieving high-precision landmark detection becomes more challenging.

This study proposes a lumbosacral localization and landmark detection network (LSLD-Net) to enhance both the accuracy of lumbar landmark detection and the precision of lumbar stability and spondylolisthesis grading assessments. To effectively utilize the prior knowledge of the lumbosacral area, fully capture its features from the global image, and eliminate the interference from extraneous information, we employ a YOLO [12] object detection network for preliminary localization of the lumbosacral area. This detection network is robust to the X-ray images with varying numbers of vertebrae and provides the accurate information for subsequent landmark detection tasks. Additionally, to achieve the high-accuracy landmark detection in the lumbosacral area, it is crucial to fully exploit the available contextual information within this region. To achieve this, we design a landmark detection network based on an enhanced U-Net [13] architecture to extract the multi-level features from the lumbar X-ray images, ranging from the low-level detail features to the high-level semantic features. This approach effectively facilitates the extraction of both the detailed and global contextual information from the images. To capture the relationship between the overall and local areas of the lumbar spine, we use the HRNet-18 [14] as the feature extractor, which allows the network to acquire both the detail-rich low-level features and the semantically rich high-level features through the frequent information exchange between the feature maps of different resolutions. Moreover, we propose a Multi-Scale Attention Module (MSAM) that emphasizes the relevant features by calculating the channel and spatial attention. This module helps the model capture the information from the local details to the global context, improving the ability of the network to learn the spatial distribution of lumbar landmarks and enhancing its robustness for accurate landmark detection. Based on these developments, the main contributions of this study are as follows:

- We propose the LSLD-Net, a network for assessing lumbar stability and grading spondylolisthesis based on lumbosacral localization and landmark detection. This network effectively integrates the multilevel semantic features from the lumbar knowledge to achieve the efficient and accurate automated landmark detection in the lumbosacral region.
- We propose a Multi-Scale Attention Module, that fully exploits the contextual information of the lumbosacral area. This module emphasizes the relevant features, such as the vertebral center areas and the lumbar corner-points, which aids the network in learning the spatial

distribution of the lumbar region and its corner-points. Simultaneously, it reduces the background and noise interference, thereby enhancing the accuracy and robustness of landmark coordinate estimation.

- We conduct the comparative experiments using both the public BUU-LSPINE dataset [15] and our private dataset, which comprises 1,398 X-ray images from various views. Our LSLD-Net achieves the highest Successful Detection Rate (SDR) in the landmark detection for both datasets, outperforming the existing methods. On our private dataset, the model achieves 70.9% accuracy in the lumbar stability assessment and 83.8% accuracy in the lumbar spondylolisthesis grading. These results demonstrate that the proposed method is effective for the clinical auxiliary diagnosis of lumbar stability and spondylolisthesis grading.

## II. METHODS

The overall framework of our proposed LSLD-Net, as illustrated in Fig. 3, comprises two main components: (a) Lumbosacral Localization Network, which employs the YOLOv8 object detection network to identify the lumbosacral area in the X-ray images and (b) Landmark Detection Network, which utilizes a fully convolutional structure with a multi-scale feature enhancement mechanism to process the lumbosacral area for accurate vertebral landmark detection.

### A. Lumbosacral Localization Network

To eliminate interference from extraneous areas on landmark detection and fully utilize the lumbosacral area information, we employ the YOLOv8 object detection network to localize the lumbosacral area (L1–L5, S1), as shown in Fig. 3 (a). The architecture consists of three main components, including Backbone, Neck, and Head. The Backbone is responsible for the feature extraction, using a series of convolutional layers to obtain the multilevel features from the input image. The Neck processes and fuses these features, incorporating a feature pyramid network to combine the feature maps from different levels, thereby improving the capture of information across various scales. The Head then predicts from the feature maps output by the Neck, with multiple prediction layers handling different scales to enhance the detection capability. Ultimately, the Head generates the bounding boxes and the classification results, accurately localizing the lumbosacral area.

Using this multi-level, the multi-scale feature fusion approach, YOLOv8 accurately localizes the lumbosacral area, demonstrating robustness against spinal X-ray images with varying numbers of vertebrae, providing accurate detection area for subsequent landmark detection tasks.

### B. Landmark Detection Network

As shown in Fig. 3 (b), our proposed landmark detection network is a variant of U-Net, featuring the specialized encoder and decoder structures to extract the semantic information of the vertebrae from the image. The vertebral semantic

features are decoupled through three independent mappings to extract the center-point heatmap, the center-point offsets, and the corner-point offsets. The landmarks are then calculated using the identified vertebral center-points and corner-point offsets, the final detection results are mapped back to the original lumbar X-ray image.

1) **Encoder:** Commonly used as a feature extractor for medical images, the U-Net has limitations, such as the loss of the high-resolution detail due to pooling operations and the single-scale semantic feature extraction via skip connections. These limitations are especially pronounced in the landmark detection for the Lumbosacral area, which requires the preservation of the high-resolution features and the multi-scale semantic information. To address this, we use an enhanced U-Net architecture with the HRNet-18 as the encoder. The HRNet-18 excels in maintaining the high-resolution features and the multi-scale feature fusion through frequent information exchange and the integration of various branches with different-resolution feature maps. This approach effectively captures both the global and local information, reflecting the overall shape of the lumbosacral region and the relationships between vertebrae, thereby significantly improving the feature extraction quality and accuracy. The final outputs in the encoder process are feature maps at the original sizes of  $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ , denoted as  $\{C1, C2, C3, C4\}$ , and the whole encoding process is as follows:

$$C1, C2, C3, C4 = F(I) \quad (1)$$

where  $I$  is the input image; and  $F(\cdot)$  is the HRNet-18.

2) **Decoder:** In each layer of the decoder, we introduce a Multi-Scale Attention Module (MSAM) to facilitate the feature fusion and selection for both the high-level and low-level features. The structure of the MSAM is illustrated in Fig. 4. First, the average pooling is used to retain the global information from the feature map, enabling the model to comprehend the overall structure of the lumbosacral area. Concurrently, the max pooling captures the important information by focusing on the most responsive pixels and highlighting key details. The results from the average and max pooling are combined using a fully connected layer with the shared parameters, and the channel attention weight  $T$  is obtained by applying the sigmoid activation function. This process is mathematically represented as follows:

$$T = \sigma(FC(P_A([F_H, F_L])) + FC(P_M([F_H, F_L]))) \quad (2)$$

where  $\sigma$  denotes the sigmoid function;  $FC(\cdot)$  denotes the fully connected layer;  $P_A(\cdot)$  denotes average pooling;  $P_M(\cdot)$  denotes max pooling; and  $[F_H, F_L]$  denotes the concatenated features of high-level features  $F_H$  and low-level features  $F_L$ .

The channel attention weight  $T$  is then element-wise multiplied by the low-level features  $F_L$  and high-level features  $F_H$ . The results are input into the Spatial Attention Enhancement (SAE) module, producing the output  $S$  of MSAM. This process is mathematically represented as follows:

$$S = [SAE(T \odot F_L), SAE(T \odot F_H)] \quad (3)$$

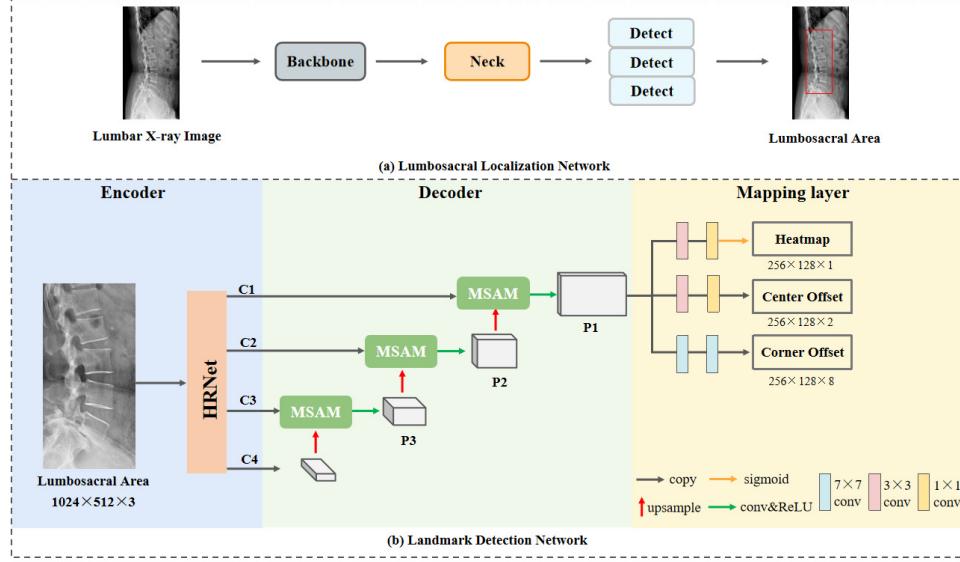


Fig. 3. Overall framework of LSLD-Net consists of two parts: (a) Lumbosacral Localization Network; (b) Landmark Detection Network.

The SAE module improves the ability of the model to focus on the important regions by directing the attention to specific parts of the image. This enhancement allows the network to concentrate on the local details of the lumbosacral area while ignoring the irrelevant background information, thus increasing the landmark detection accuracy. The SAE module redistributes the attention weights along the spatial dimension, enabling the model to capture the detailed landmark features more effectively. In the SAE, the channel max pooling and average pooling are first applied along the spatial dimension. The results are concatenated and processed through a series of convolution operations, followed by the sigmoid function to derive the spatial attention weights. These weights are then multiply element-wise by the feature  $x$ . This process is mathematically represented as follows:

$$SAE(x) = \sigma(Conv([C_A(x), C_M(x)])) \odot x \quad (4)$$

where  $C_A(\cdot)$  denotes the channel average pooling operation;  $C_M(\cdot)$  denotes the channel max pooling operation;  $Conv(\cdot)$  denotes a series of convolution operations; and  $\odot$  denotes an element-wise multiplication.

The MSAM effectively captures both the local details and the global contextual information in the lumbosacral region by computing channel and spatial attention, thereby significantly enhancing the landmark detection performance. By emphasizing the relevant features, such as the lumbar center and the lumbar corner-points, the MSAM assists the network in learning the spatial distribution of the lumbar vertebrae and their landmarks. Simultaneously, it minimizes the impacts of background noise and irrelevant features, improves the fusion of the high-level and low-level features, and enhances the robustness.

**3) Mapping Layer:** In the lumbosacral X-ray images, similar vertebral anatomy and complex backgrounds can inhibit

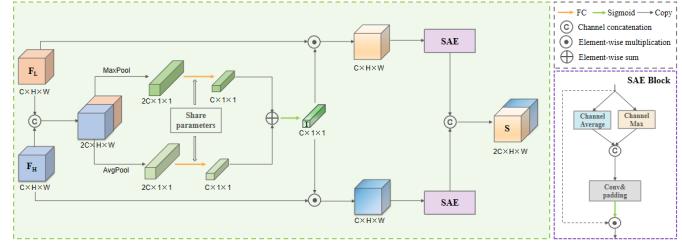


Fig. 4. A block diagram of Multi-Scale Attention Module

the direct regression of landmark coordinates, leading to the inaccurate localization. To address this, we follow the landmark detection strategy from a previous work [11]. First, we localize each vertebra using a heatmap of the vertebral center-points and correct any quantification errors using the center-point offsets. Subsequently, we compute the corner coordinates based on the determined vertebral center-points and corner offsets. This approach allows us to initially focus on the global structure of the vertebrae with the center-point heatmap and subsequently capture the detailed features with the offsets, resulting in more accurate landmark detection.

In this method, the mapping layer has three prediction targets: the center-point heatmap, the center-point offsets, and the corner offsets, which together determine the landmarks. The output  $P1$  of the decoder includes the semantic information regarding the position and shape of each vertebra, which is then transformed into three distinct feature spaces by the mapping layer. This process is mathematically represented as follows:

$$\text{Heatmap} = \sigma(Conv_{1\times 1}(Conv_{3\times 3}(P1))) \quad (5)$$

$$\text{Center Offset} = Conv_{1\times 1}(Conv_{3\times 3}(P1)) \quad (6)$$

$$\text{Corner Offset} = Conv_{7\times 7}(Conv_{7\times 7}(P1)) \quad (7)$$

where  $\text{Conv}_{n \times n}(\cdot)$  denotes the convolution kernel is an  $n \times n$  convolution operation.

**4) Loss:** The overall loss function in this study contains three components, including the center-point heatmap, the center-point offset, and the corner offset. Because the center-point coordinates are used as a reference for predicting all the subsequent corner-point coordinates, their accuracy is crucial for the landmark detection results of the four corner-points. Consequently, different weights are assigned to each component. The vertebral landmark detection loss function can be mathematically represented as follows:

$$\text{Loss} = \alpha_1 L_{hm} + \alpha_2 L_{reg} + \alpha_3 L_{wh} \quad (8)$$

where  $L_{hm}$ ,  $L_{reg}$ , and  $L_{wh}$  are the losses calculated by the network for the center-point heatmap, center-point offset and corner offset, respectively. Here,  $\alpha_2 = 5$ ,  $\alpha_1$  and  $\alpha_3$  are 1.

**Center-Point Heatmap:** The center-point  $k$  of each vertebrae is used to generate a single-channel center-point ground truth heatmap with an unnormalized Gaussian kernel, as shown in Fig. 5 (b). The loss for the center-point is optimized using the focal loss.

**Center-Point Offset:** The center-point offset is utilized to correct the quantization errors introduced by downsampling. Specifically, the center-point position  $(x, y)$  of a vertebra in the input image is mapped onto the downsampled  $n$ -fold feature map as  $(\lfloor \frac{x}{n} \rfloor, \lfloor \frac{y}{n} \rfloor)$  and the center-point offset is defined as  $(\frac{x}{n} - \lfloor \frac{x}{n} \rfloor, \frac{y}{n} - \lfloor \frac{y}{n} \rfloor)$ . The loss associates with the center-point offset is trained using the L1 loss.

**Corner Offset:** As shown in Fig. 5 (c), the corner offset is represented as the vectors pointing from the center-point to the four corner-points of the vertebrae, denoted as  $(\frac{x_c}{n} - \frac{x_c}{n}, \frac{y_c}{n} - \frac{y_c}{n})$ , where  $(\frac{x_c}{n}, \frac{y_c}{n})$  are the coordinates of the vertebral corner-points. The loss associated with corner offsets is trained using the L1 loss.

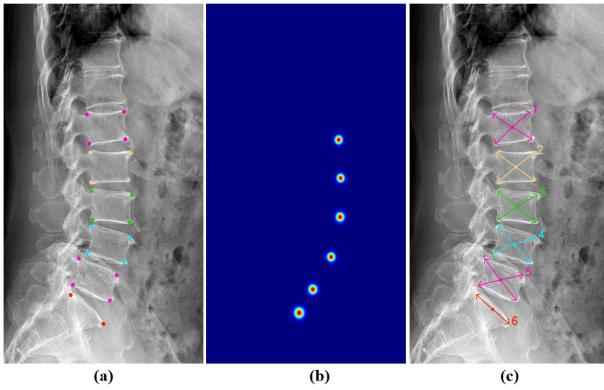


Fig. 5. Landmarks, center-point heatmap, and corner offset: (a) each X-ray image shows a total of 22 landmarks for five lumbar vertebrae and one sacral vertebrae; (b) center-point heatmap; (c) corner offset.

### III. EXPERIMENTS AND RESULTS

#### A. Dataset

This study uses two datasets to validate the proposed method, both of which are relevant to the lumbar stability

and spondylolisthesis grading tasks. These datasets consist of a public dataset and a private dataset.

The public dataset consists of 400 lateral X-ray images provided by BUU-LSPINE [15]. These images are annotated by professional doctors, with each lumbar vertebra labeled at four corner-points and the sacrum labeled at its upper two corner-points, resulting in 22 landmarks, as shown in Fig. 5 (a). Of these images, 350 are adopted for the training and validation, whereas 50 are reserved for testing.

To advance the study, we collect the X-ray images from 466 patients at a local hospital, encompassing both healthy and diseased patients. Each patient provides the hyperextension, lateral, and hyperflexion X-ray images, resulting in a total of 1,398 X-ray images. The images are in the PNG format with a pixel spacing of 0.143 mm. An experienced radiologist and orthopedic surgeon annotates and aligns the dataset collaboratively, following the same annotation process as the public dataset. From this collection, 379 cases with 1,137 X-ray images are applied for the training and validation, whereas 87 cases with 261 X-ray images are reserved for testing.

#### B. Experimental Settings

Our method is implemented in PyTorch and executes on a computer with an NVIDIA GeForce RTX 4090 (24 GB of memory). To mitigate overfitting, the data augmentation techniques, such as random flipping, random scaling, and contrast variation, are applied. The original images are cropped to 1024×512 pixels before being input into the network, with a batch size of 2. The Adam optimizer is employed for the network optimization, with an initial learning rate of  $1.25 \times 10^{-4}$ , and the network is trained for a total of 100 epochs.

#### C. Evaluation Metrics

First, we evaluate the reliability of the detection results on both the public and private datasets using the standard metrics in the landmark detection tasks: Mean Absolute Error (MAE), Mean Radial Error (MRE), and Successful Detection Rate (SDR). Lower MAE and MRE values indicate better performance, whereas higher SDR values are preferred. The specific calculation formulae are mathematically represented as follows:

$$\text{MAE} = \frac{1}{M} \frac{1}{K} \sum_{m=1}^M \sum_{n=1}^K |gt_n^l - pred_n^l| \quad (9)$$

$$\text{MRE} = \frac{1}{M} \frac{1}{K} \sum_{m=1}^M \sum_{n=1}^K \|gt_n^l - pred_n^l\|_2 \quad (10)$$

where  $M$  denotes the number of samples;  $K$  denotes the number of landmarks per X-ray image;  $gt_n^l$  and  $pred_n^l$  are the true and predicted coordinates of the landmarks.

SDR measures the rate of landmarks successfully detected, where a landmark is deemed successfully detected if MRE between the predicted and true coordinates is within a specified error range  $\delta$ . For the public dataset, the error range is set to 5–10 px, and for the private dataset, it is set to 2–5 mm,

TABLE I  
COMPARISON OF PROPOSED AND EXISTING METHODS ON PUBLIC AND PRIVATE DATASETS. THE BEST RESULTS ARE IN **BOLD** TEXT AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	Public Dataset							Private Dataset						
	MAE (px)	MRE (px)	SDR(%)					MAE (mm)	MRE (mm)	SDR(%)				
			5px	6px	7px	8px	9px			2mm	3mm	4mm		
Hourglass-2	11.0	8.8	59.5	69.3	75.6	80.9	83.5	86.2	5.0	4.0	64.8	81.7	88.9	92.0
Hourglass-8	10.8	8.6	66.0	73.8	78.9	83.1	85.6	87.9	4.5	3.6	70.7	85.6	91.8	93.8
HRNet	11.8	9.5	75.7	80.6	83.0	85.4	87.0	87.8	4.0	3.3	80.4	90.5	94.0	95.1
Yi et al. [11]	10.2	8.2	<u>77.0</u>	81.1	84.4	86.3	88.1	89.0	<u>3.8</u>	<u>3.1</u>	<u>81.1</u>	91.1	94.1	95.2
Ao et al. [17]	<u>10.1</u>	<u>8.0</u>	76.1	<b>82.9</b>	<u>85.4</u>	<u>87.1</u>	<u>88.4</u>	<u>89.8</u>	3.9	3.2	80.2	<u>91.3</u>	<u>94.5</u>	<u>95.7</u>
<b>Ours</b>	<b>8.9</b>	<b>7.1</b>	<b>78.1</b>	<u>82.7</u>	<b>86.6</b>	<b>88.8</b>	<b>90.3</b>	<b>90.9</b>	<b>3.5</b>	<b>2.9</b>	<b>81.9</b>	<b>92.2</b>	<b>95.1</b>	<b>96.2</b>

based on the image pixel spacing of 0.143 mm. The specific calculation formula is mathematically represented as follows:

$$SDR_{\delta} = \frac{\text{Count}\left(\left\{\text{pred}_n^l : \|\text{gt}_n^l - \text{pred}_n^l\|_2 \leq \delta\right\}\right)}{\text{Count}(\Omega_{\text{test}})} \quad (11)$$

where  $\text{Count}(\cdot)$  is a counting function that counts the number of landmarks that meet the  $(\cdot)$ ;  $\Omega_{\text{test}}$  are all landmarks in the test set.

Second, to evaluate the clinical feasibility of the model, we use the following metrics on a private dataset: the mean error of the ISA (Error<sub>ISA</sub>), the mean error of vertebral SD (Error<sub>SD</sub>), the accuracy of lumbar stability assessment (ACC<sub>d</sub>), and the accuracy of lumbar spondylolisthesis grading (ACC<sub>s</sub>). The specific calculation formulae are mathematically represented as follows:

$$\text{Error}_{\text{ISA}} = \frac{1}{M} \frac{1}{5} \sum_{m=1}^M \sum_{n=1}^5 |\text{ISA}_n^{gt} - \text{ISA}_n^{\text{pred}}| \quad (12)$$

$$\text{Error}_{\text{SD}} = \frac{1}{M} \frac{1}{5} \sum_{m=1}^M \sum_{n=1}^5 |\text{SD}_n^{gt} - \text{SD}_n^{\text{pred}}| \quad (13)$$

where  $\text{ISA}_n^{gt}$  and  $\text{ISA}_n^{\text{pred}}$  denote the true and predicted values of the change in lumbar ISA corresponding to the hyperextension and hyperflexion views, respectively. Similarly,  $\text{SD}_n^{gt}$  and  $\text{SD}_n^{\text{pred}}$  represent the true and predicted vertebral SD changes for these views.

Based on the criteria for assessing the lumbar stability and the spondylolisthesis grading outlined in this study, the ACC<sub>d</sub> and the ACC<sub>s</sub> calculation formulae are mathematically represented as follows:

$$\text{ACC}_d = \frac{C_d}{T} \quad (14)$$

$$\text{ACC}_s = \frac{C_s}{T} \quad (15)$$

where  $C_d$  and  $C_s$  denote the number of vertebrae correctly predicted for the lumbar spondylolisthesis grading and the lumbar stability, respectively; T denotes the total number of vertebrae across all the samples.

#### D. Comparative experiments

To evaluate the efficacy of the proposed second-stage landmark detection network, we compare it with both classical and advanced vertebral landmark detection methods using the public and private datasets. All the experimental inputs are the lumbosacral areas processed using the first-stage object-detection network, and the final landmark detection results from the second stage are then mapped back to the original lumbar X-ray images. The comparison results are presented in Table I. The Hourglass network [16] and High Resolution Network [14] are the classical methods for the landmark detection, with the Hourglass-2 and Hourglass-8 representing the models stack with two and eight hourglass modules, respectively. Yi et al. [11] and Ao et al. [17] exemplify the leading approaches in spine landmark detection.

The results indicate that our proposed LSND-Net achieves the lowest MAE and MRE on both the public and private datasets. Specifically, on the public dataset, the MAE and MRE are 8.9 and 7.1 px, respectively, whereas on the private dataset, they are 3.5 and 2.9 mm, respectively. For the SDR, our model reaches the highest values across all error ranges on the private dataset, notably achieving 95% at a 4 mm error range. On the public dataset, our model also performs the best in most error ranges, whereas it demonstrates the slight underperformance at 6 px compared to Ao et al. However, it significantly outperforms other methods in all other error ranges, including achieving the SDR above 90% at 9 and 10 px error ranges. These results demonstrate that our model is robust, comprehensive, and achieves the high landmark detection accuracy across different datasets.

In addition, we perform a qualitative analysis of the landmark detection results for the methods discussed using both private and public datasets, as depicted in the Fig. 6 and Fig. 7. Our model demonstrates a higher degree of overlap between the red and green points (examples highlighted by the yellow rectangular box), which is further supported by the lower MRE. In Fig. 6, due to variations in vertebral morphology across different views of the patient, some methods exhibit significant errors in landmark detection. Specifically, in the hyperextension view shown in Fig. 6(c), both the Hourglass

and Yi et al. methods demonstrate a substantial misalignment in landmark detection (highlighted by the blue rectangular area). This misalignment may result from insufficient capture of contextual and detailed information in the lumbosacral area, leading to inaccurate vertebral center-point localization. In contrast, our model consistently achieves high landmark detection accuracy across different patient views of X-ray images. Because the public dataset only contains lateral views, in Fig. 7, we select landmark detection results from three different patients to further demonstrate the reliability of our model. These visualization results illustrate that our model maintains high accuracy across different views and datasets, demonstrating excellent robustness.

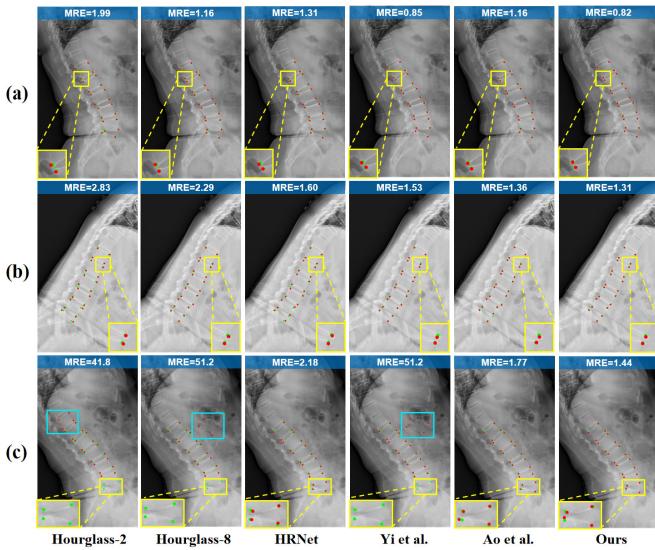


Fig. 6. Qualitative comparison of different methods on our private dataset. The green points represent the ground-truth landmarks, and the red points are the predicted results. For reference, the MRE value is displayed at the top of the image. (a), (b), and (c) show landmark detection for the same patient in lateral, hyperflexion, and hyperextension views, respectively.

To demonstrate the clinical significance of LSND-Net's high-precision landmark detection, the comparative experiments are conducted on a private dataset to assess its effectiveness in evaluating the lumbar stability and the spondylolisthesis grading. The results presented in Table II indicate that our model achieves the lowest mean errors in ISA and vertebral SD. In the spondylolisthesis grading, our model achieves the highest accuracy of 83.8%, while its accuracy in lumbar stability assessment exceeds 70%, significantly surpassing other models. Although Ao et al. and Yi et al. achieve better landmark detection metrics than HRNet, their clinical assessment accuracies for lumbar stability and spondylolisthesis grading are lower. This discrepancy may result from the minor errors in landmark detection affecting the angle measurements and the vertebral SD, which mislead the classification at the judgment threshold, thus affecting the clinical assessment results.

These experiments assess the landmark detection accuracy and clinical assessment performance of LSND-Net, confirming its robustness and comprehensiveness from various perspectives.

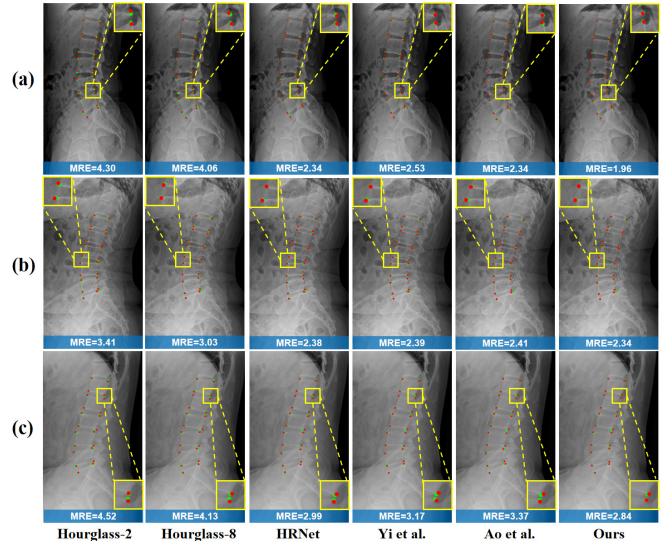


Fig. 7. Qualitative comparison of different methods on a public dataset. The green points represent the ground-truth landmarks, and the red points are the predicted results. The MRE value is displayed at the bottom of the image. (a), (b), and (c) show landmark detection for three patients in lateral views.

TABLE II  
COMPARISON OF PROPOSED AND EXISTING METHODS ON PRIVATE DATASETS FOR LUMBAR SPINE STABILITY ASSESSMENT AND SPONDYLOLISTHESIS GRADING. THE BEST RESULTS ARE IN **BOLD**.

Method	Error <sub>ISA</sub> (°)	Error <sub>SD</sub> (mm)	ACC <sub>d</sub> (%)	ACC <sub>s</sub> (%)
Hourglass-2	3.9	2.0	79.5	64.2
Hourglass-8	3.4	1.9	82.3	65.6
HRNet	<b>2.2</b>	1.7	82.7	69.9
Yi et al. [11]	2.4	1.5	82.0	67.8
Ao et al. [17]	2.5	1.6	82.1	69.8
<b>Ours</b>	<b>2.2</b>	<b>1.4</b>	<b>83.8</b>	<b>70.9</b>

The LSND-Net significantly enhances the accuracy of lumbar imaging parameters, thereby aiding the clinicians in the early quantitative diagnosis and evaluation of lumbar spinal stability and spondylolisthesis grading.

#### E. Ablation experiments

To evaluate the effectiveness of each module in our model, we conduct the ablation experiments on both the public and private datasets comprising four experimental groups. The results, detailed in Table III, are as follows. Base refers to replacing the traditional U-Net encoder with the HRNet structure. The second row highlights the impact of incorporating the first-stage lumbosacral object detection module in LSND-Net, which enhances the focus on the lumbosacral region and minimizes the interference from irrelevant areas. This addition results in a reduction of the MRE by 0.9 px on the public dataset and 0.8 mm on the private dataset. The third row illustrates the effect of including the MSAM without the SAE module, demonstrating the improved detection accuracy due to channel attention enhancement. The fourth row indicates that

TABLE III

ABLATION EXPERIMENTS OF PROPOSED METHOD ON PUBLIC AND PRIVATE DATASET WITH YOLO, MSAM, SAE. THE BEST RESULTS ARE IN **BOLD**.

Method	Public Dataset							Private Dataset						
	MAE (px)	MRE (px)	SDR(%)					MAE (mm)	MRE (mm)	SDR(%)				
			5px	6px	7px	8px	9px			2mm	3mm	4mm	5mm	
Base	14.0	11.4	70.7	77.5	81.0	83.6	85.2	85.9	4.8	4.0	75.2	86.8	90.6	92.1
Base+YOLO	13.0	10.5	73.9	78.4	81.2	83.5	85.0	86.1	3.9	3.2	80.9	90.8	93.7	95.0
Base+YOLO+MSAM(w/o SAE)	10.5	8.4	74.0	81.1	84.5	85.5	86.5	88.5	3.7	3.1	80.2	91.1	94.4	95.7
<b>Base+YOLO+MSAM(w/ SAE)</b>	<b>8.9</b>	<b>7.1</b>	<b>78.1</b>	<b>82.7</b>	<b>86.6</b>	<b>88.8</b>	<b>90.3</b>	<b>90.9</b>	<b>3.5</b>	<b>2.9</b>	<b>81.9</b>	<b>92.2</b>	<b>95.1</b>	<b>96.2</b>

adding the SAE module to MSAM further enhances the ability to capture detailed landmark features, leading to an improved SDR. The final model achieves a SDR of 90.9% within a 10 px error range on the public dataset and 96.2% within a 5 mm error range on the private dataset.

The experiments confirm the necessity and effectiveness of the various modules in our model, demonstrating their role in enhancing the accuracy of the landmark detection network.

#### IV. SUMMARY

In this study, we propose a novel assessment network based on Lumbosacral Localization and Landmark Detection (LSDL-Net), which provides the high-precision landmark estimation for the lumbosacral area and enhances the clinical assessment of the lumbar vertebral stability and the spondylolisthesis grading. Initially, we employ YOLOv8 for the accurate localization of the lumbosacral area in the X-ray images, which minimizes the interference from the extraneous information such as thoracic vertebrae and reduces the landmark misalignment. We then integrate the advanced HRNet and U-Net architectures and introduce a Multi-Scale Attention Module, which captures the long-distance contextual information, the overall structural layout, and the fine local details in the lumbar spine X-ray images, leading to the higher-precision landmark detection. The evaluation of both the private and public BUU-LSPINE datasets demonstrates that our LSDL-Net model surpasses other state-of-the-art methods in the landmark detection, significantly improving the accuracy and efficiency of the lumbar stability assessment and the spondylolisthesis grading, thereby aiding the clinicians in early quantitative diagnosis and assessment.

#### ACKNOWLEDGMENT

This research work was supported by the Ningbo Municipal Public Welfare Technology Research Project (No.2022S134) and the Ningbo Major Research and Development Plan Program (Grant No. 2023Z196).

#### REFERENCES

- [1] Berven, Sigurd, and Rishi Wadhwa. "Sagittal Alignment of the Lumbar Spine." Neurosurgery clinics of North America 29.3 (2018): 331-339.
- [2] Hu, Houmin, et al. "Development and validation of an automatic diagnostic tool for lumbar stability based on deep learning." Chinese Journal of Reparative and Reconstructive Surgery 37.1 (2023): 81-90.
- [3] Meyerding, Henry W. "Spondylolisthesis." JBJS 13.1 (1931): 39-48.
- [4] Zhang, Junhua, et al. "Computer-aided cobb measurement based on automatic detection of vertebral slopes using deep neural network." International journal of biomedical imaging 2017.1 (2017): 9083916.
- [5] Smith, Justin S., et al. "Treatment of adult thoracolumbar spinal deformity: past, present, and future: JNSPG 75th Anniversary Invited Review Article." Journal of Neurosurgery: Spine 30.5 (2019): 551-567.
- [6] Reddy, Pavan Kumar, et al. "Anatomical landmark detection using deep appearance-context network." 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021.
- [7] Nguyen, Thong Phi, et al. "Deep learning system for Meyerding classification and segmental motion measurement in diagnosis of lumbar spondylolisthesis." Biomedical Signal Processing and Control 65 (2021): 102371.
- [8] Zhang, Kailai, et al. "MPF-net: An effective framework for automated cobb angle estimation." Medical Image Analysis 75 (2022): 102277.
- [9] Liu, J., C. Yuan, and X. Sun. "The measurement of Cobb angle based on spine X-ray images using multi-scale convolutional neural network. Phys Eng Sci Med 44: 809-821." (2021).
- [10] Li, Xiang, et al. "FDGR-Net: Feature Decouple and Gated Recalibration Network for medical image landmark detection." Expert Systems with Applications 238 (2024): 121746.
- [11] Yi, Jingru, et al. "Vertebra-focused landmark detection for scoliosis assessment." 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.
- [12] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [13] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.
- [14] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- [15] Klinwichit, Podchara, et al. "BUU-LSPINE: A thai open lumbar spine dataset for spondylolisthesis detection." Applied Sciences 13.15 (2023): 8646.
- [16] Newell, Alejandro, K. Yang, and J. Deng. "Stacked Hourglass Networks for Human Pose Estimation." European conference on computer vision 2016.
- [17] Ao, Yueyuan, and Hong Wu. "Feature aggregation and refinement network for 2D anatomical landmark detection." Journal of Digital Imaging 36.2 (2023): 547-561.