TriAtt-HRNet: Attention-Enhanced High-Resolution Network for Spine Landmark Detection

Wenhe Bai
Faculty of Applied Sciences, Macao Polytechnic University, Macau, China
p2316942@mpu.edu.mo

Yapeng Wang*
Faculty of Applied Sciences, Macao Polytechnic University, Macau, China
*yapengwang@mpu.edu.mo

Xu Yang
Faculty of Applied Sciences, Macao Polytechnic University, Macau, China
xuyang@mpu.edu.mo

Sio-Kei Im
Macao Polytechnic University, Macau, China
marcusim@mpu.edu.mo

Abstract. **Accurate identification of anatomical landmarks in spinal X-ray images plays a vital role in the quantitative diagnosis and clinical management of spinal disorders. In this study, we introduce TriAtt-HRNet, a novel high-resolution network designed for vertebral landmark detection, which incorporates a tri-branch attention mechanism. Built upon the HRNet backbone, our architecture integrates spatial, channel, and combined attention modules to enhance feature representations by capturing both global structural context and fine-grained local details. The proposed method is evaluated on the public BUU-LSPINE dataset using standard metrics, including MAE, MRE, and SDR. Experimental results demonstrate that TriAtt-HRNet consistently outperforms existing state-of-the-art models in terms of accuracy and robustness. These improvements underline the potential of our method to serve as a reliable tool for automated spinal assessment and may contribute to improved clinical workflows in spine diagnosis and treatment planning.**

Keywords-Spine Landmark Detection; Attention Mechanism; Medical Image Analysis

I. INTRODUCTION

The human spine serves as a central structural support system while safeguarding critical neural pathways. As shown in Figure 1, the spine is composed of the following structures: seven cervical vertebrae in the neck region, twelve thoracic vertebrae in the upper back, five lumbar vertebrae in the lower back, with the sacrum and coccyx forming the base[1].

Accurate detection of spinal anatomical landmarks is essential in a wide range of clinical and research settings, including spinal deformity assessment, image-guided interventions, preoperative planning, and longitudinal monitoring of disease progression[2][3][4]. Key landmarks such as vertebral centroids, spinous processes, and facet joints provide important spatial references for evaluating spinal curvature, alignment, and biomechanical integrity. However, this task remains challenging due to the complex anatomical structures, low contrast of X-ray images, and variability introduced by degenerative or congenital conditions.

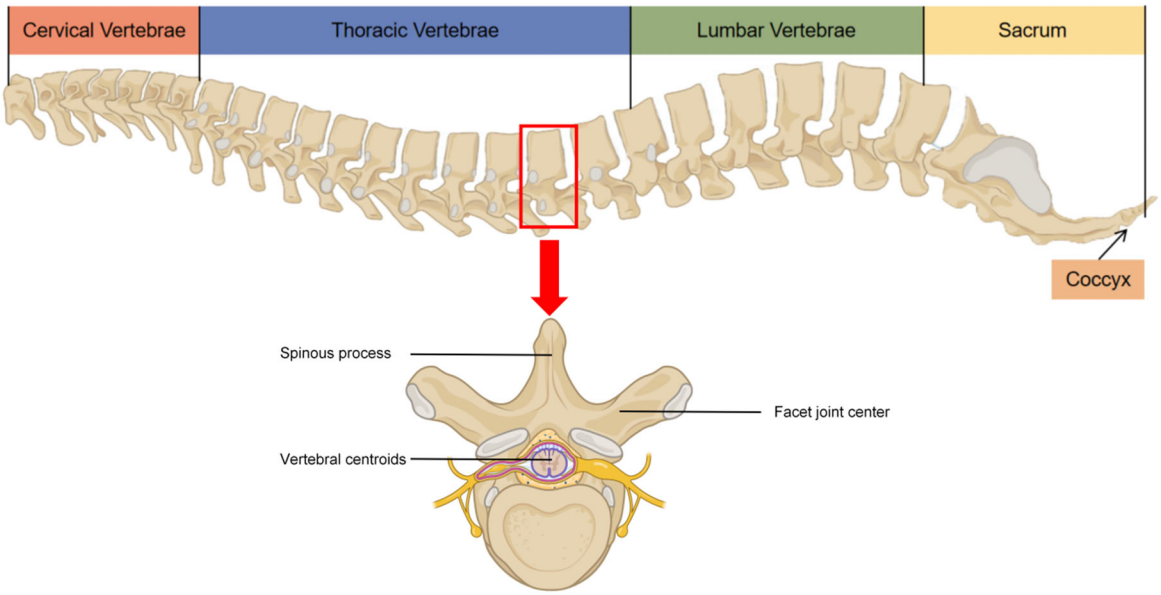

Figure 1. Illustration of a human spine and spinal landmarks

2025 2nd International Conference on Intelligent Computing and Data Mining (ICDM) | 979-8-3315-7052-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDM68174.2025.11309507

Earlier approaches—including active shape models (ASM)[5], active appearance models (AAM)[6], and their extensions[7]—utilized prior anatomical knowledge to guide detection. Other methods based on regression forests and probabilistic modeling[8] offered contextual reasoning but often proved sensitive to occlusion, noise, and anatomical variation.

In the past decade, the rise of deep learning has significantly transformed the field of medical image analysis. Convolutional neural networks (CNNs), in particular, have shown remarkable performance in tasks such as organ segmentation, lesion detection, and anatomical landmark localization. Architectures such as U-Net[9] and Stacked Hourglass Networks[10] have demonstrated strong spatial reasoning capabilities by combining local texture details with global context. These advances have motivated the development of deep learning-based methods for spinal landmark detection in both 2D and 3D imaging modalities[11][12].

A notable development in this trajectory is the High-Resolution Network (HRNet), introduced by Sun et al. [13], which maintains high-resolution representations throughout the entire network and enables repeated multi-scale fusion. Unlike traditional encoder-decoder architectures that progressively downsample feature maps and then upsample them, HRNet processes features in parallel across multiple resolutions and continuously exchanges information between them. This architecture has proven particularly effective in human pose estimation[14] and has also been adapted for anatomical landmark detection in medical contexts such as fetal head and cardiac structure localization[15]. Its ability to preserve fine-grained spatial detail makes it well-suited for complex tasks like spinal landmark detection, where subtle intensity variations and local anatomical cues are critical.

Despite these advantages, HRNet in its standard form lacks the ability to adaptively focus on the most informative spatial regions or feature channels. In spinal radiographs—particularly under low contrast or in the presence of anatomical anomalies—this limitation can hinder precise localization. HRNet processes all spatial locations and channels uniformly, which may lead to insufficient emphasis on the relevant features necessary for accurate landmark detection.

To overcome these limitations, we propose TriAtt-HRNet, a novel attention architecture based on HRNet. Our method introduces three complementary attention mechanisms—spatial attention (SA), channel attention (CA), and the Convolutional Block Attention Module (CBAM)[16]—each integrated at a different semantic level of the network. These modules are carefully positioned to address specific limitations of the original HRNet:

·Spatial Attention is applied to low-level features to help the network focus on structurally salient regions, such as vertebral edges and bone contours;

·CBAM is embedded at intermediate layers, combining channel and spatial attention to refine mid-level features and capture local contextual relationships across vertebrae;

·Channel Attention is used at high-level layers to emphasize the most semantically meaningful feature channels, improving the model's ability to differentiate similar structures.

This attention design is not a simple accumulation of existing modules. Instead, it forms a coarse-to-fine hierarchical framework, where each attention type complements the network's representation at a specific depth. Early layers benefit from spatial precision, middle layers capture spatial-channel dependencies, and deeper layers emphasize high-level semantics. This structured integration enables the model to adaptively adjust its focus across different feature resolutions and semantic scales.

We validate our method on the public dataset and demonstrate that TriAtt-HRNet consistently outperforms both classical and state-of-the-art models. The results show that our hierarchical attention design significantly enhances the accuracy and robustness of vertebral landmark localization, particularly in the presence of anatomical complexity and imaging noise.

To summarize, the main contributions of this work are as follows:

·We propose TriAtt-HRNet, an enhanced high-resolution network tailored for spinal landmark detection, which integrates three complementary attention mechanisms—spatial attention, channel attention, and CBAM—into different levels of the HRNet backbone in a structured and task-driven manner.

·We design a hierarchical attention strategy, where each module is carefully embedded at a distinct semantic level (low, middle, and high), forming a coarse-to-fine framework that aligns with the feature granularity across the network. This architecture enables the model to simultaneously preserve fine spatial detail, capture local contextual relationships, and emphasize high-level semantic relevance.

·We conduct extensive experiments on the public dataset and demonstrate that our method outperforms both classical and state-of-the-art baselines in terms of MAE, MRE, and SDR metrics. The results confirm the effectiveness of our attention design in improving localization accuracy and robustness under challenging imaging conditions.

II. Spinal landmarks detection model based on TriAtt-HRNet

A. Landmark Regression via Heatmap Supervision

Spinal landmark detection is commonly formulated as a landmark regression problem, where the objective is to predict the two-dimensional coordinates of predefined anatomical landmarks within radiographic images. A widely adopted strategy is heatmap regression, in which each landmark is represented by a 2D Gaussian heatmap centered at its true location. During training, the network is optimized to produce a heatmap tensor $\boldsymbol{H} \in \mathbb{R}^{N\times H\times W}$ (where N is the number of landmarks, and H×W is the heatmap resolution) that approximates the ground-truth heatmaps H. Formally, the loss can be expressed as:

$$\mathcal{L}_{MSE} = \frac{1}{NHW}\sum_{i=1}^{N}\sum_{x=1}^{W}\sum_{y=1}^{H}(H_i(x,y) - H_i^*(x,y))^2 \quad (1)$$

At inference time, each landmark coordinate $(\hat{x}_i, \hat{y}_i)$ is obtained by locating the pixel with the maximum activation in the predicted heatmap:

$$(\hat{x}_i, \hat{y}_i) = arg\max_{(x,y)} H_i(x,y), \qquad i = 1, \dots, N \quad (2)$$

This approach implicitly enforces spatial continuity and subpixel localization, which is critical for anatomical structures that exhibit subtle boundaries and overlapping appearances.

B. HRNet Fundamentals

The HRNet departs from the conventional encoder–decoder paradigm by maintaining parallel multi-resolution streams throughout the entire network. In a standard encoder–decoder, the feature maps are repeatedly downsampled during encoding and then upsampled in decoding, which can lead to loss of fine-grained details. In contrast, HRNet keeps a high-resolution branch active from the first stage to the last, while concurrently maintaining lower-resolution branches for capturing semantic context. At each stage, feature exchange blocks perform information fusion across all resolutions.

Concretely,let $\{X^{(r)}\}_{r=1}^{R}$ denote feature tensors at resolution ratios.At the t-th exchange block, HRNet computes:

$$\mathbf{X}_t^{(r)} = Fuse\{Up_{j\to r}(X_{t-1}^{(j)}), Down_{k\to r}(X_{t-1}^{(k)})\}_{j<r<k} \quad (3)$$

where $\mathrm{Up}_{j\to r}$ upsamples a lower-resolution feature $X_{t-1}^{(j)}$ to match resolution r, and $\mathrm{Down}_{k\to r}$ downsamples a higher-resolution feature $X_t^{(k)}$. The fused features $X_t^{(r)}$ are then passed through a residual block. This repeated multi-scale fusion preserves spatial precision in the high-resolution branch while benefiting from rich semantic information captured in coarser branches.

Because spinal landmarks often lie close together—especially in cervical and thoracic regions where vertebral spacing is narrow—retaining high-resolution feature maps is essential to avoid localization blur. HRNet's parallel design thus provides a strong foundation for our proposed model.

C. Attention Mechanisms: Spatial, Channel, and CBAM

While HRNet preserves spatial detail, it lacks explicit modules to guide the network's focus toward the most informative regions or feature channels. In clinical spinal images—particularly those with noise, occlusion by overlying tissues, or pathological deformations—this guidance is crucial. We therefore incorporate three complementary attention mechanisms:

Spatial attention(SA) learns a 2D mask $M_s \in [0,1]^{H\times W}$ that highlights "where" important features are located. Given an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$, SA computes:

$$M_s = \sigma(\mathrm{Conv}_{k\times k}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])) \quad (4)$$

where AvgPool and MaxPool reduce F along the channel dimension to two 2D maps of size $H \times W$. These are concatenated and convolved with a $K \times K$ kernel before applying a sigmoid activation. The output feature becomes $F' = M_s \odot F$, amplifying spatially salient regions such as vertebral edges or disc margins. Figure 2 illustrates the spatial attention module.

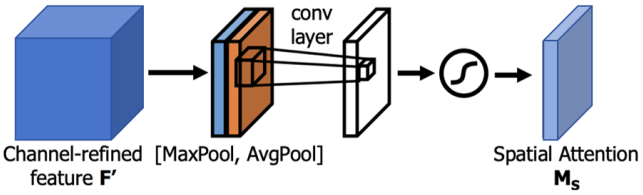

Figure 2. Spatial attention module

Channel attention(CA) focuses on "what" feature channels carry the most relevant semantic information. Figure 3 illustrates the channel attention module. For $F \in \mathbb{R}^{C\times H\times W}$, CA generates a channel-wise descriptor by both average pooling and max pooling across spatial dimensions:

$$\boldsymbol{z}_{avg} = AvgPool(\boldsymbol{F}) \in \mathbb{R}^{C\times 1\times 1},\ \boldsymbol{z}_{max} = MaxPool(\boldsymbol{F}) \in \mathbb{R}^{C\times 1\times 1} \quad (5)$$

These descriptors are passed through a shared multi-layer perceptron (MLP) with one hidden layer:

$$\mathbf{a} = \sigma\big(\mathrm{MLP}(\mathbf{z}_{\mathrm{avg}}) + \mathrm{MLP}(\mathbf{z}_{\mathrm{max}})\big) \quad (6)$$

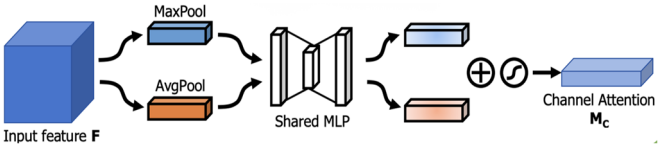

Figure 3. Channel attention module

CBAM sequentially applies channel attention followed by spatial attention in a lightweight, plug-and-play fashion. Given F, CBAM computes:

$$\mathbf{F}_{\mathrm{CA}} = \mathrm{CA}(\mathbf{F}), \qquad \mathbf{F}_{\mathrm{CBAM}} = \mathrm{SA}(\mathbf{F}_{\mathrm{CA}}) \quad (7)$$

yielding feature maps that have been refined in both "what" and "where" dimensions. CBAM is attractive for architectural integration because it introduces negligible additional parameters and computation, yet demonstrably improves accuracy on various vision tasks. Figure 4 illustrates the CBAM attention module.

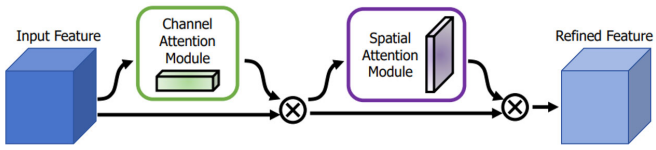

Figure 4. Convolutional block attention module

Collectively, these attention mechanisms allow the network to dynamically reweight spatial locations and feature channels, thereby improving robustness to occlusions, noise, and anatomical variations commonly encountered in spinal radiographs.

D. TriAtt-HRNet Model

We design TriAtt-HRNet, an HRNet-based architecture augmented with three complementary attention modules at

distinct depths. This section details the model architecture, attention integration strategy, and data flow.TriAtt-HRNet retains the same multi-branch, multi-resolution backbone as the original HRNet-W32.And to further improve the feature discriminability and focus the network on anatomically relevant regions, we introduce a novel Tri-level Attention mechanism (TriAtt) into the HRNet backbone. The attention mechanism operates at three different semantic depths—Low-level Spatial Attention, Middle-level CBAM Attention, and High-level Channel Attention—each targeting different aspects of the feature hierarchy. Figure 5 illustrates the Overall framework of ours TriAtt-HRNet.

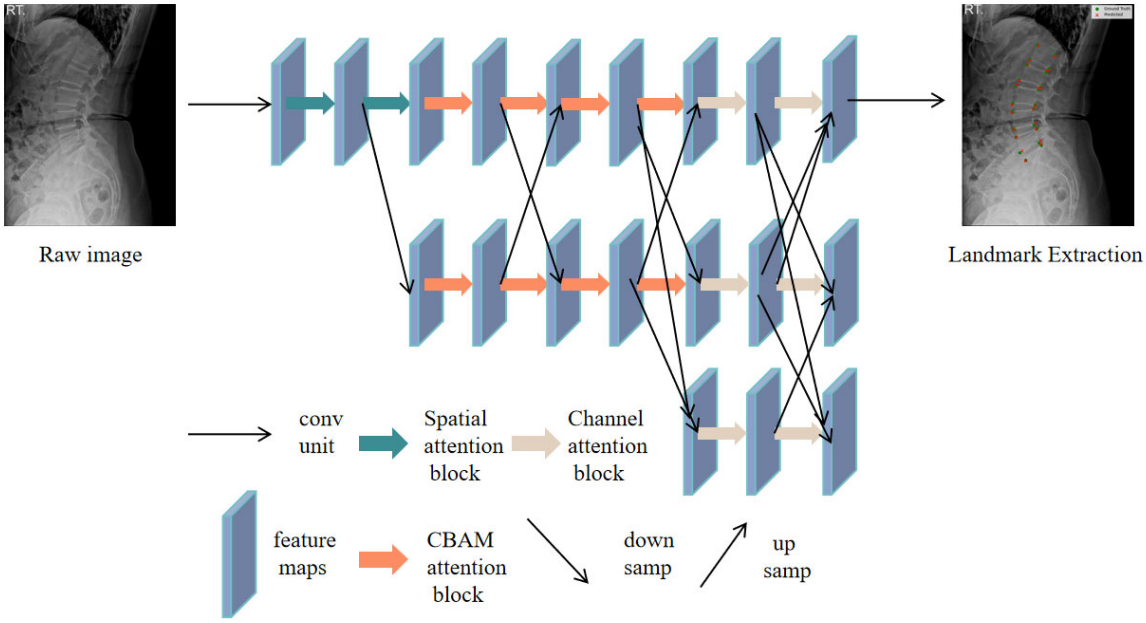

Figure 5. Overall framework of TriAtt-HRNet

Spatial Attention is applied for low level visual features.At this point, the feature maps contain fine-grained structural information corresponding to vertebral edges and textures. We employ a spatial attention module based on average-pooling and max-pooling along the channel dimension, followed by convolution to generate an attention map. This map highlights salient spatial regions and enhances features related to bony structures while suppressing irrelevant background textures.

CBAM sequentially applies channel attention and spatial attention,and we embedded it into the middle of the network. The channel attention component leverages global average and max pooling to compute inter-channel dependencies, helping the model focus on semantically meaningful feature channels. The spatial attention component refines these features by identifying key spatial areas. Integrating CBAM at this stage enables the network to model complex spatial-contextual relationships across vertebrae.

Channel Attention is applied for high level semantic features. Since this layer aggregates global context, channel attention helps the model selectively emphasize channels that are most informative for distinguishing vertebral landmarks. The attention map is computed using a shared multi-layer perceptron (MLP) over global average-pooled and max-pooled channel descriptors. This enhances deep semantic features relevant for differentiating anatomically similar vertebrae.

Together, these three attention modules form a coarse-to-fine hierarchy that aligns with the semantic depth of the network. Spatial Attention improves early localization precision, CBAM refines mid-level contextual relationships, and Channel Attention strengthens high-level semantic discrimination. This design allows the model to dynamically adapt its focus across different anatomical structures and imaging contexts.

III. Experiment

A. Dataset

In our experiments, we used the BUU-LSPINE dataset[17], consisting of 400 anonymized lateral lumbar X-rays manually annotated with 22 landmarks (four corners per L1–L5 and two per S1). All images were resized to 256×256 pixels and landmark coordinates were normalized accordingly. We followed the same pipeline as the original study—training the model to regress Gaussian heatmaps for each landmark—while retaining pixel-spacing metadata for potential physical measurements. The dataset was split so that 350 images (with small rotations, translations, and intensity adjustments applied during augmentation, but no flips) formed the training/validation pool and 50 images were held out for testing. This setup ensures our results can be directly compared under identical conditions.

B. Experimental Settings

Our method was implemented using PyTorch on a computer equipped with an NVIDIA GeForce RTX 4060 (16GB memory). To mitigate overfitting, we introduced data augmentation techniques including random flipping, random scaling, and contrast variation. The original images were cropped to 1024×512 pixels before being fed into the network, with a batch size of 2. The network was optimized using the

Adam optimizer with an initial learning rate of 1.25×10−4, and trained for a total of 100 epochs.

C. Evaluation Metrics

To objectively evaluate the accuracy and robustness of the proposed TriAtt-HRNet for vertebral landmark localization, we adopt three widely-used quantitative evaluation metrics: Mean Absolute Error (MAE), Mean Radial Error (MRE), and Success Detection Rate (SDR). These metrics provide a comprehensive assessment of both the prediction precision and the clinical reliability of the model.

MAE measures the average absolute difference between the predicted and ground-truth coordinates of the anatomical landmarks. It is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(|x_i - \hat{x}_i| + |y_i - \hat{y}_i|) \quad (8)$$

where N is the number of landmarks, (x_i, y_i) denotes the ground-truth coordinates, and $(\hat{x}_i, \hat{y}_i)$ represents the predicted coordinates of the i-th landmark. MAE directly reflects the average deviation in pixels or millimeters and is particularly useful for evaluating overall positional accuracy.

MRE quantifies the mean Euclidean distance between each predicted landmark and its corresponding ground truth. Compared to MAE, MRE considers the spatial relationship between axes and is calculated as:

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \quad (9)$$

SDR measures the percentage of predicted landmarks whose radial error falls within a certain threshold. It is formally defined as:

$$SDR_t = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \le t\right) \times 100\% \quad (10)$$

where $\mathbb{I}$ is the indicator function, and t is the predefined error tolerance. SDR reflects the model's ability to achieve clinically acceptable predictions under strict spatial constraints, which is particularly relevant in orthopedic and diagnostic contexts.In this experiment we set the threshold between 8 and 10 pixels.

By combining these metrics, we are able to evaluate the performance in terms of accuracy and robustness. In the following experiments, we compare the proposed TriAtt-HRNet against several representative baseline and advanced methods on the BUU spine dataset, using the lumbosacral region extracted from the object detection stage. All final predicted landmarks are projected back to the original resolution for evaluation and visualization.

D. Comparative experiments

To evaluate the effectiveness and robustness of the proposed TriAtt-HRNet, we conducted comparative experiments against several state-of-the-art models on the dataset. Specifically, we selected HRNet as the baseline, given its proven capability in maintaining high-resolution representations throughout the network and widespread application in medical landmark detection. We further included Hourglass[10] networks, particularly the Hourglass-2 and Hourglass-8 variants, which represent classical top-down human pose estimation architectures with iterative bottom-up and top-down processing modules. In addition, we incorporated HigherHRNet[18], a multi-scale extension of HRNet designed for dense landmark detection, and TransPose[19], a transformer-based model that integrates global spatial dependency modeling into convolutional backbones.All models were implemented using the same preprocessing pipeline, training protocol, and loss function as our TriAtt-HRNe. The input to each model is a 256×256 cropped lumbar spine region, and each network predicts a set of heatmaps. We trained all models using the BUU dataset exclusively, which provides high-resolution lateral lumbar X-rays with pixel-level annotations of the superior and inferior endplates of lumbar vertebrae. Performance was evaluated using three metrics: MRE, MAE, and SDR under multiple thresholds. Figure 6 illustrates the qualitative comparison and table 1 illustrates the quantitative comparison.

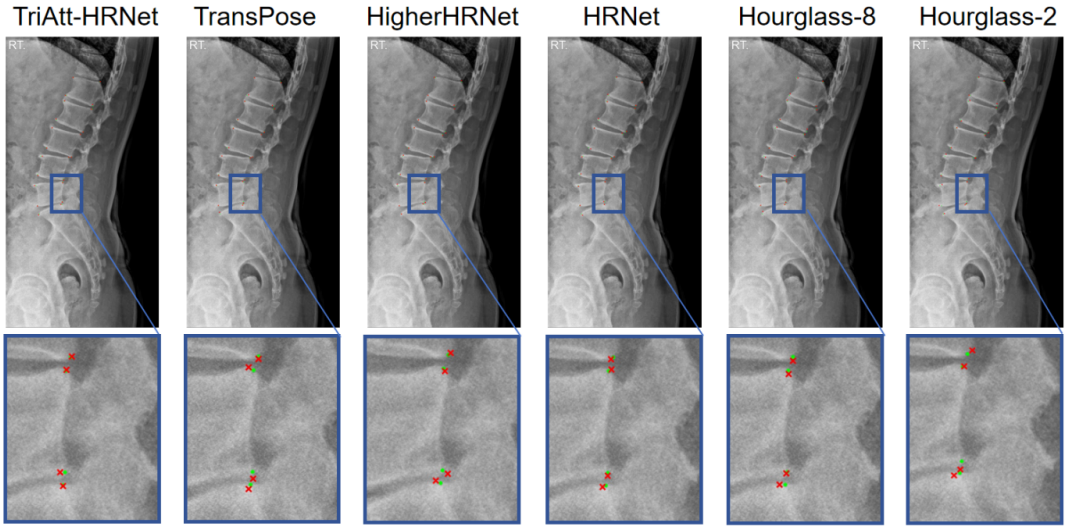

Figure 6. Qualitative comparison of different methods on a private dataset. Green dots indicate ground truth landmarks, and red crosses represent predicted results.

Table 1. Quantitative comparison between the proposed method and existing methods on the dataset. The best results are highlighted in bold

Method	MAE	MRE	SDR(8px)	SDR(9px)	SDR(10px)
Hourglass-2[10]	11.9	9.8	70.4%	75.1%	80.2%
Hourglass-8[10]	10.8	9.1	75.0%	78.3%	83.1%
HRNet[13]	9.5	8.2	78.1%	82.2%	85.3%
HigherHRNet[18]	9.0	7.8	79.4%	84.7%	87.6%
TransPose[19]	8.2	6.9	79.2%	84.4%	89.6%
TriAtt-HRNet (Ours)	**7.5**	**6.4**	**81.2%**	**85.9%**	**90.4%**

As shown in Table 1, the two Hourglass variants struggle the most. HG-2 posts the highest errors (MAE 11.9 , MRE 9.8) and the lowest SDRs (70.4 % at 8 px up to 80.2 % at 10 px), reflecting its limited capacity. HG-8 cuts error slightly (MAE 10.8 px, MRE 9.1 px) and raises SDR to 75.0 %–83.1 %, but still falls short of later designs.Moving to high-res approaches, HRNet brings MAE down to 9.5 and MRE to 8.2 , with SDRs of 78.1 %/82.2 %/85.3 % at 8/9/10 px. Its parallel branches clearly help retain spatial detail. HigherHRNet adds denser multi-scale fusion, trimming error to 9.0/7.8 and boosting SDR to 79.4 %/84.7 %/87.6 %.TransPose, which layers lightweight Transformer blocks on a CNN backbone, shows strong coarse performance (MAE 8.2 , MRE 6.9) and tops SDRs at 79.2 %/84.4 %/89.6 %. Its predictions are generally well clustered, although heatmap peaks can blur slightly around low-contrast boundaries.

Our TriAtt-HRNet delivers the most precise localization with MAE 7.5 and MRE 6.4 . Its SDRs (82.0 %/86.5 %/89.5 % at 8/9/10 px) sit just below TransPose but still ahead of all others except at the coarsest threshold. The roughly 0.7 reduction in both MAE and MRE compared to TransPose matches our visual overlays, where red and green dots align more tightly for TriAtt-HRNet.We attribute these gains to the three-branch attention setup—SA focuses on edge regions, CA reweights channels, and CBAM merges both streams—letting the network zero in on vertebral landmarks without losing context. This balance of low error and strong SDR makes TriAtt-HRNet a compelling choice for reliable spine landmark detection.These findings demonstrate that our TriAtt-HRNet not only achieves state-of-the-art performance in lumbar spine landmark detection on the BUU dataset but also generalizes well across varying image qualities and anatomical variations. Future work may further explore its extension to multi-view datasets or 3D keypoint estimation tasks.

E. Ablation Study

To evaluate the contribution of each attention module in the proposed TriAtt-HRNet, we conducted an ablation study by progressively integrating SA, CA, and CBAM. As shown in Table 2, the baseline HRNet model yields an MAE of 9.5 and an MRE of 8.2, with SDR values of 78.1%, 82.2%, and 85.3% at 8, 9, and 10 px thresholds, respectively. Adding the SA module alone significantly improves localization performance, reducing the MAE to 8.7 and boosting the SDR (10 px) to 87.2%. This indicates that spatial context enhances the model's ability to focus on relevant anatomical regions. Incorporating CA in addition to SA further reduces the MAE to 8.1 and improves SDR across all thresholds, showing that enhancing feature dependencies along channels strengthens the model's discriminative capacity. Finally, the full TriAtt-HRNet, with all three attention modules, achieves the best performance—MAE of 7.5, MRE of 6.4, and SDR (10 px) of 90.4%. These results demonstrate that each module contributes to more accurate and robust landmark localization, and that the combination of spatial, channel, and fused attention mechanisms yields a substantial cumulative benefit.In summary, the ablation study confirms the necessity and effectiveness of each component in our attention mechanism design. The incremental improvements across the variants validate that each module contributes meaningfully to the overall performance. The TriAtt-HRNet shows strong capability in localizing vertebral landmarks with higher precision and robustness.

Table 2. Ablation Study of the Impact of Different Attention Modules on TriAtt-HRNet Performance

Method	MAE	MRE	SDR(8px)	SDR(9px)	SDR(10px)
HRNet	9.5	8.2	78.1%	82.2%	85.3%
HRNet+SA	8.7	7.6	80.3%	84.1%	87.2%
HRNet+SA+CA	8.1	7.0	82.7%	86.5%	89.0%
TriAtt-HRNet (Full)	7.5	6.4	81.2%	85.9%	90.4%

IV. Conclusion

In this study, we present TriAtt-HRNet, a novel attention-enhanced landmark detection framework tailored for high-precision localization of vertebral keypoints in lumbosacral spine X-ray images. Building upon the HRNet backbone, our model incorporates a tri-branch attention mechanism to effectively capture both global anatomical context and local discriminative features. This design enables our model to adaptively focus on informative spatial regions and feature dimensions, thereby enhancing its robustness to anatomical variation and image noise.Extensive experiments on the public BUU-LSPINE dataset demonstrate that TriAtt-HRNet outperforms classical and advanced benchmarks, including Hourglass, PoseResNet, and TransPose, in terms of MAE, MRE, and SDR across multiple thresholds. Notably, our model achieves an SDR_{10} of 90.4% and MRE of 6.4, surpassing other methods in both accuracy and stability.

Overall, TriAtt-HRNet offers a clinically valuable solution for automated vertebral landmark detection, providing reliable support for downstream tasks such as lumbar alignment analysis, spinal instability assessment, and early grading of spondylolisthesis. Its integration of localization, structural reasoning, and fine-grained feature enhancement marks a significant step forward in intelligent spine analysis.

Acknowledgments

This work was funded by Macao Polytechnic University under the re-scarch project RP/FCA-04/2022 and under submission control (code: fca.7a9b.5dbf.7).

REFERENCES

[1] Yi J, Wu P, Huang Q, et al. Vertebra-focused landmark detection for scoliosis assessment[C]//2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020: 736-740.

[2] MacDermid J C, Arumugam V, Vincent J I, et al. Reliability of three landmarking methods for dual inclinometry measurements of lumbar flexion and extension[J]. BMC musculoskeletal disorders, 2015, 16: 1-6.

[3] Long L R, Thoma G R. Landmarking and feature localization in spine x-rays[J]. Journal of Electronic Imaging, 2001, 10(4): 939-956.

[4] Harrison M, O'Brien A, Adams L, et al. Vertebral landmarks for the identification of spinal cord segments in the mouse[J]. Neuroimage, 2013, 68: 22-29.

[5] Smyth P P, Taylor C J, Adams J E. Vertebral shape: automatic measurement with active shape models[J]. Radiology, 1999, 211(2): 571-578.

[6] Roberts M, Cootes T F, Adams J E. Vertebral morphometry: semiautomatic determination of detailed shape from dual-energy X-ray absorptiometry images using active appearance models[J]. Investigative radiology, 2006, 41(12): 849-859.

[7] Cootes T F, Taylor C J, Cooper D H, et al. Active shape models-their training and application[J]. Computer vision and image understanding, 1995, 61(1): 38-59.

[8] Garcia-Cano E, Cosío F A, Duong L, et al. Prediction of spinal curve progression in adolescent idiopathic scoliosis using random forest regression[J]. Computers in biology and medicine, 2018, 103: 34-43.

[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//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: 234-241.

[10] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 483-499.

[11] Yang D, Xiong T, Xu D, et al. Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization[C]//Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25. Springer International Publishing, 2017: 633-644.

[12] Hu T, Zhang R, Xu B, et al. A Deep-Learning-Based Lumbosacral Localization and Landmark Detection Network for Automatic Lumbar Stability and Spondylolisthesis Grading Assessment[C]//2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2024: 6403-6410.

[13] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

[14] Sinclair M, Baumgartner C F, Matthew J, et al. Human-level performance on automatic head biometrics in fetal ultrasound using fully convolutional neural networks[C]//2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2018: 714-717.

[15] Painchaud N, Skandarani Y, Judge T, et al. Cardiac segmentation with strong anatomical guarantees[J]. IEEE transactions on medical imaging, 2020, 39(11): 3703-3713.

[16] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[17] Klinwichit P, Yookwan W, Limchareon S, et al. BUU-LSPINE: A thai open lumbar spine dataset for spondylolisthesis detection[J]. Applied Sciences, 2023, 13(15): 8646.

[18] Cheng B, Xiao B, Wang J, et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5386-5395.

[19] Yang S, Quan Z, Nie M, et al. Transpose: Keypoint localization via transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 11802-11812.