RouterNet: Hierarchical Point Routing Network for Robust Vertebral Landmark Localization on AP X-ray Images

Yingjie Guo[1*], Jinxin Lv[1*], Wei Fang[2], Qiang Li[1†], Zhiwei Wang[1†]

[1]Huazhong University of Science and Technology
[2]Wuhan United Imaging Surgical Healthcare Co., Ltd.
guoyingjie@hust.edu.cn, lvjinxin@vivo.com, fenghui@hzau.edu.cn, liqiang8@hust.edu.cn, zwwang@hust.edu.cn

Abstract

Locating vertebral landmarks on anteroposterior (AP) X-ray images is challenging due to the tissue overlap. Despite the great progress of heatmap-based methods, they often predict missing/false points, which are intolerable in the downstream applications like scoliosis assessment. In this paper, we instead modernize the classic point-regression scheme, and propose a novel model termed RouterNet to locate the 68 vertebral landmarks completely and accurately. RouterNet starts from an initial root point, and then gradually routes it onto more and more points with finer and finer semantics. RouterNet naturally couples such point routing process with its hierarchical and multi-scale feature learning. That is, lower-scale feature maps are utilized to regress points with coarser semantics, and the regressed points pilot a more focused local feature extraction on the next higher-scale map to route onto their subsequent positions with finer semantics. With this divide-and-conquer, RouterNet alleviates the task difficulty, and can robustly localize by routing from the whole spinal center to 17 vertebral centers, and further to their 68 corner points. Extensive and comprehensive experiments on both public and private datasets demonstrate our superior performance over other state-of-the-arts, by decreasing NMSE by 73.8% for landmark localization, and SMAPE by 14.8% for the downstream scoliosis assessment.

Code — https://github.com/YingJGuo/RouterNet

Introduction

Scoliosis is characterized by lateral spinal curvature accompanied by vertebral rotation (Wang et al. 2021). Manual Cobb angle measurement, the current clinical gold standard, is time-consuming and suffers from high inter-observer variation. Accurately locating vertebral landmarks is crucial for automatic computer-aided scoliosis diagnosis and downstream applications like Cobb angle calculation.

X-ray imaging is widely used in scoliosis diagnosis due to its availability, economy, and lower radiation dose compared to CT. An anteroposterior (AP) X-ray image contains 12 thoracic and 5 lumbar vertebrae, with landmarks defined as four corner points of each vertebra (68 total). However, as shown in Fig. 1(a), tissue superposition makes the spine low-contrast, and vertebra-like structures cause confusion, making automatic localization extremely challenging.

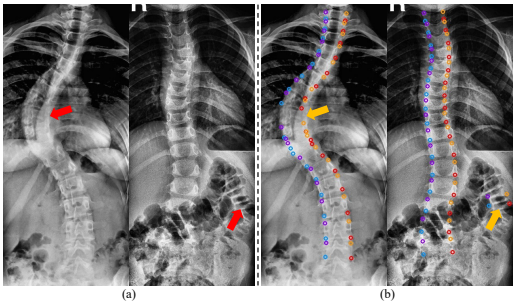

Figure 1: (a): Tissue overlap and vertebra-like structures in AP X-ray images. (b): Missing/false points often predicted by heatmap-based methods. The different colors encode four corner points.

Current solutions divide into two categories: heatmap-based and point-regression methods.

Heatmap-based methods generate Gaussian heatmaps for each landmark and locate points via local maxima, offering robustness to pose variations, and thus have dominated in tasks of natural scenes. However, the great robustness to large poses usually comes with a price of two types of errors, i.e., missing points (see Fig. 1(b) left) induced by dilution of overlapped tissues, and false points (see Fig. 1(b) right) incurred by other vertebra-like structures. Such errors result in dramatically misguiding biomarkers, and cause misdiagnosis in subsequent clinical applications.

Therefore, we in this paper argue that *the heatmap-based methods are overkill in our focusing task, since the robustness to large poses is of secondary importance whereas a correct shape topology (no missing/false, only slightly misaligning) is the primary.* Although there are several efforts (Payer et al. 2019; Wang et al. 2022) that have laboriously tried to embed a prior shape into the heatmap generation, the other classical point regression scheme can retain the shape topology effortlessly.

Point-regression methods initialize points and update them progressively, naturally preserving shape topology. However, the highly nonlinear mapping from images to

[*]These authors contributed equally.
[†]Corresponding authors

landmarks makes direct regression challenging. Recent work (Wang et al. 2024) proposed multi-stage cascaded CNNs, but this complicates training and creates dependency on first-stage predictions.

In view of the above analysis, we in this paper propose a novel and efficient CNN model termed RouterNet, which absorbs the idea of cascade point regression to learn the complex mapping, yet by cascading the intrinsic decoding layers instead of separate models. The task difficulty is further alleviated via a hierarchical point routing strategy on the top of decoded multi-scale global feature maps. Specifically, using the global maps one by one in increasing order of scale, RouterNet first routes the input points onto more points with finer semantics, and then further adjust their positions for the next routing. RouterNet employs a differentiable Precision RoI Pooling (Jiang et al. 2018) to select more focused local features conditioning on the guidance of previously localized points, and also permits itself to be trained in an end-to-end manner as a whole.

In summary, our contributions are listed as follows:

- We propose RouterNet, a novel and efficient CNN model for robust vertebral landmark localization by coupling its in-built multi-scale layers with a classic cascaded point regression process.
- We alleviate the learning difficulty by hierarchically routing points from coarse to fine in terms of both quantity and semantics. Point-guided local feature selection in RouterNet also permits an end-to-end training and usage of more discriminative visual cues for predicting points with finer semantics.
- Extensive and comprehensive experiments on two public and one private datasets demonstrate the superior performance of RouterNet in both the focusing task of landmark localization and the downstream task of scoliosis assessment. Furthermore, RouterNet can almost eliminate missing/false points, and thus shows great potential for other clinical applications.

Related Work

Heatmap-based Landmark Localization

Heatmap-based approaches have achieved success in computer vision (Yu and Tao 2021; Sun et al. 2019) and been applied to vertebral landmarks. FARNet (Ao and Wu 2023) predicts 68 individual heatmaps while Zhang *et al.* (Zhang et al. 2021) use 6 grouped semantic heatmaps. Both suffer from quantization errors due to discrete arg-max operations (Tompson et al. 2015). Yi *et al.* (Yi et al. 2020) and Guo *et al.* (Guo et al. 2021b) address this via coordinate offset regression and Transformers respectively. However, heatmap methods struggle with missing detections and false positives due to tissue occlusion and vertebra-like structures, limiting clinical applicability.

Point-regression Landmark Localization

Point-regression methods directly update initialized points toward targets, preserving shape topology. Classical approaches include Active Appearance Models (Matthews and Baker 2004) and Explicit Shape Regression (Cao et al. 2014). For vertebral landmarks, Sun *et al.* (Sun et al. 2017) used Structured Support Vector Regression while Wu *et al.* (Wu et al. 2017) proposed BoostNet. Although these methods eliminate missing/false points, single-stage regression struggles with the complex nonlinear mapping from visual features to landmark coordinates. Wang *et al.* (Wang et al. 2024) addressed this via cascaded CNNs with PCA constraints, but stage-by-stage training complicates optimization and creates dependency chains. Our RouterNet unifies cascading within a single end-to-end network using hierarchical point routing.

Method

Revisiting Classic Cascaded Point Regression

The key insight of cascaded point regression is using several regressors to gradually fit a set of initialized points to the target positions stage by stage. In each stage, the regressor uses shape-indexed features to estimate the coordinate offsets. Given an input image I, the process of cascaded point regression can be formulaically expressed as follows:

$$\hat{S}^t = \hat{S}^{t-1} + Reg^t(F_{index}(I, \hat{S}^{t-1})), t = 1, \ldots, T \quad (1)$$

where T is the number of total stages, $\hat{S}^t$ is the estimated coordinates of points in the t^{th} stage, F_{index} is the shape-indexed feature extraction conditioning on points, and Reg^t is the learnable regressor for the t^{th} stage. The initial $\hat{S}^0$ is often the average of all training labels, $\hat{S}^0 = 1/N \sum S_n$, where S_n is the n^{th} sample's ground-truth (GT) coordinates. In the training phase, the regression target of Reg^t is the offsets between the GT and previous-stage coordinates of points, that is, $\theta^t_{Reg} = \arg\min 1/N \sum \|Reg^t_n - (S_n - \hat{S}^{t-1}_n)\|_2$.

There are two key issues when translating the above cascaded point regression using the concept of deep neural networks. First, the complication of training is unbearable if each Reg^t is an independent CNN model. Second, the optimization is separate for each stage since F_{index} is none-differentiable. To address these, our proposed RouterNet reinvents the cascaded process as a coarse-to-fine hierarchical point routing, and tightly couples it with the CNN's inherent multi-scale feature learning. RouterNet is also equipped with a differentiable local feature selection guided by points, making itself enjoy an end-to-end optimization as a whole.

RouterNet: CNN-style Point Regression

We denote a training sample as $\{I, (S_{rt}, S_{ctr}, S_{cnr})\}$, where I is the AP X-ray image, and $S_{cnr} \in \mathbb{R}^{68\times2}$ is the manually annotated coordinates of the entire 68 vertebral landmarks. $S_{ctr} \in \mathbb{R}^{17\times2}$ is the GT 17 vertebrae's center points, which is the average of every 4 corners, and S_{rt} is simply defined as one of the centers. We set S_{rt} to the 9^{th} center point, which is the closest point to the whole spine center.

These root, center and corner points (S_{rt}, S_{ctr}, S_{cnr}) are hierarchical representations of the same spine, but with more

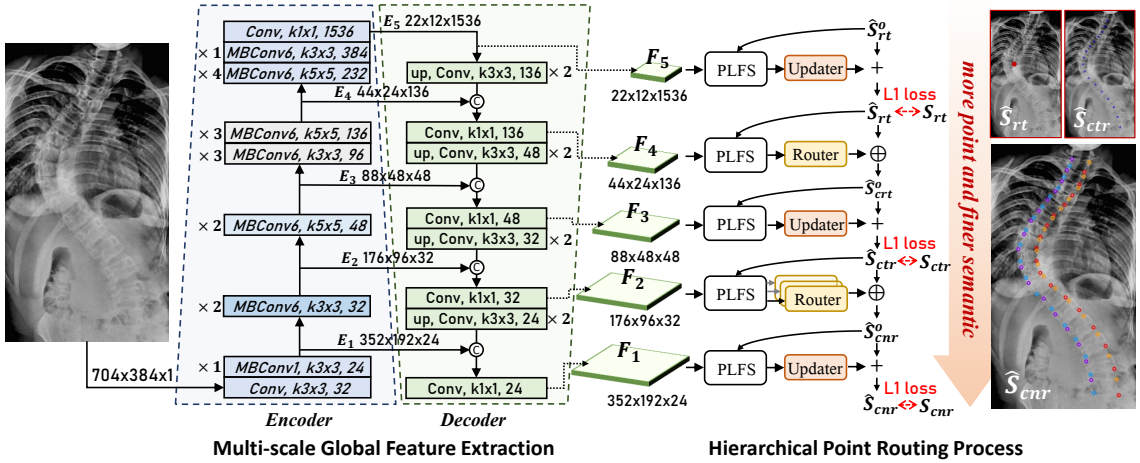

Figure 2: RouterNet contains two parts, i.e., Multi-scale Global Feature Extraction and Hierarchical Point Routing Process. Point-guided Local Feature Selection (PLFS) bridges the two parts as well as routers/updaters, enabling an end-to-end training of RouterNet as a whole.

and more points as well as finer and finer semantics. The relationship between them is formulated as follows:

$$\begin{cases} S_{cnr} = \{(x_{cnr}, y_{cnr})_k\}, \\ S_{ctr} = \{(x_{ctr}, y_{ctr})_k\} = \{\frac{1}{4}\sum_{i=1}^{4}(x_{cnr}, y_{cnr})_{4(k-1)+i}\}, \\ S_{rt} = (x_{rt}, y_{rt}) = (x_{ctr}, y_{ctr})_9, \end{cases} \quad (2)$$

where $k = 1, \ldots, 68$ for S_{cnr} and $k = 1, \ldots, 17$ for S_{ctr}.

RouterNet follows a coarse-to-fine hierarchical routing path: $\hat{S}_{rt} \rightarrow \hat{S}_{ctr} \rightarrow \hat{S}_{cnr}$. This coarse-to-fine path can be naturally coupled with the inherent multi-scale feature learning in CNN. Motivated by this, RouterNet adopts the framework as shown in Fig. 2, consisting of two major parts, i.e., Multi-scale Global Feature Extraction and Hierarchical Point Routing Process.

Multi-scale Global Feature Extraction RouterNet resizes input I to 704×384 and uses a modified EfficientNet-B3 (Tan and Le 2019) encoder-decoder to extract 5 multi-scale feature maps $\{\boldsymbol{F_1}, \ldots, \boldsymbol{F_5}\}$ with downsampling ratios $2^2, 4^2, 8^2, 16^2, 32^2$. The encoder consists of MBConv6 blocks with squeeze-and-excitation optimization (Tan et al. 2019; Hu, Shen, and Sun 2018). The decoder progressively fuses encoding features$\{\boldsymbol{E_1}, \ldots, \boldsymbol{E_5}\}$ with lower-scale maps to recover semantic information.

Hierarchical Point Routing Process Similar to Eq. (1), RouterNet also starts from an initialized root point. RouterNet makes the feature extraction and point regression highly entangled with each other. The hierarchical point routing process can be formulated as follows:

$$\begin{cases} \hat{S}^o_{rt} = \frac{1}{N}\sum(S_{rt})_n, \\ \hat{S}_{rt} = \hat{S}^o_{rt} + \mathtt{U_{rt}}(\mathtt{PLFS}(\boldsymbol{F_5}, \hat{S}^o_{rt})), \\ \hat{S}^o_{ctr} = \hat{S}_{rt} \oplus \mathtt{R_{rt\rightarrow ctr}}(\mathtt{PLFS}(\boldsymbol{F_4}, \hat{S}_{rt})), \\ \hat{S}_{ctr} = \hat{S}^o_{ctr} + \mathtt{U_{ctr}}(\mathtt{PLFS}(\boldsymbol{F_3}, \hat{S}^o_{ctr})), \\ \hat{S}^o_{cnr} = \hat{S}_{ctr} \oplus \mathtt{R_{ctr\rightarrow cnr}}(\mathtt{PLFS}(\boldsymbol{F_2}, \hat{S}_{ctr})), \\ \hat{S}_{cnr} = \hat{S}^o_{cnr} + \mathtt{U_{cnr}}(\mathtt{PLFS}(\boldsymbol{F_1}, \hat{S}^o_{cnr})) \end{cases} \quad (3)$$

where $(S_{rt})_n$ is the n^{th} sample's GT root, superscript o means a primary status for routing, $\mathtt{U}(\cdot)$ and $\mathtt{R}(\cdot)$ are updater and router for estimating coordinate offsets, $\oplus$ means broadcast-adding, and $\mathtt{PLFS}(\cdot)$ is Point-guided Local Feature Selection (PLFS) which will be detailed in the following.

RouterNet uses lightweight MLP-based routers and updaters that efficiently select local features from static decoding pools. The differentiable PLFS enables end-to-end optimization across the entire path in Eq. (3).

PLFS: Point-guided Local Feature Selection

The shape-indexed feature plays a key role in the cascaded shape regression, whereas the previous implementations lack differentiability, impeding a joint learning of both multi-scale feature extractor and point regressors.

In this case, we propose a differentiable parametric module termed PLFS as detailed in Fig. 3, which extracts local features $\{\boldsymbol{z}_k\}$ from a global map $\boldsymbol{F}$ conditioning on the given points $S = \{(x, y)_k\} \in \mathbb{R}^{K\times 2}$, where the k^{th} point $(x, y)_k$ corresponds to the local feature vector $\boldsymbol{z}_k$ (see Eq. (4)).

$$\{\boldsymbol{z}_k\} = \mathtt{PLFS}\,(\boldsymbol{F}, S; w, h) \in \mathbb{R}^{Kd} \quad (4)$$

where d is channel length identical to $\boldsymbol{F}$, and (w, h) defines a local window of interest as indicated in Fig. 3.

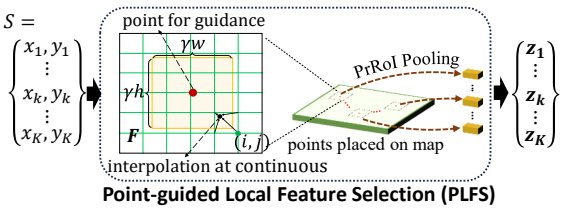

Figure 3: Details in Point-guided Local Feature Selection (PLFS) module. Given points S and a global feature map $\boldsymbol{F}$, PLFS generates a set of local feature vectors, each of which $\boldsymbol{z}_k$ is an differentiable integration within an interesting local windows.

Essentially, $\texttt{PLFS}(\cdot)$ consists of multiple Precise RoI-Pooling layers (Jiang et al. 2018), which ensures gradient back propagation. Each layer corresponds to a point in S, yielding a local feature vector from $\boldsymbol{F}$. Specifically, a continuous map $\boldsymbol{f}$ (the yellow region in Fig. 3) has to be defined using the following equation:

$$\boldsymbol{f}(x,y) = \sum_{i,j} IC\left(x,y,i,j\right) \times \boldsymbol{F}(i,j) \tag{5}$$

where $IC(x,y,i,j) = \max(0,1-|x-i|) \times \max(0,1-|y-j|)$ is an interpolation coefficient, (i,j) is a discrete index of $\boldsymbol{F}$, and (x,y) is a continuous index of $\boldsymbol{f}$. Therefore, $\boldsymbol{z}_k$ is calculated by integrating over $\boldsymbol{f}$ within the scaled local window centering at the k^{th} point, as formulated in Eq. (6).

$$\boldsymbol{z}_k = \frac{\int_{\gamma(y_k-h/2)}^{\gamma(y_k+h/2)} \int_{\gamma(x_k-w/2)}^{\gamma(x_k+w/2)} \boldsymbol{f}(x,y) dxdy}{\gamma^2 \times w \times h} \tag{6}$$

where k indexes the point in S, γ is a downsampling factor between the feature map $\boldsymbol{F}$ and input image I, that is, $\gamma = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\right\}$ for $\{\boldsymbol{F_1}, \ldots, \boldsymbol{F_5}\}$, and the window size (w,h) is set to $(96,64)$ in our case.

Point Updating and Routing

The entire updating and routing path in Eq. (3) can be viewed as a chain of connected segments. Each segment is a sub-path $S^o \xrightarrow[\texttt{U}]{\boldsymbol{F}} S \xrightarrow[\texttt{R}]{\boldsymbol{F}_{\texttt{next}}} S^o_{\texttt{next}}$ (see Fig. 4), where $S^o \in \mathbb{R}^{K\times 2}$ is from the initialization or the previous time of routing. $S^o_{\texttt{next}} \in \mathbb{R}^{KM\times 2}$ is the routed points, M means the number of finer points corresponding to each point in S.

Specifically, RouterNet first concatenates the differentiable shape-indexed local features into a single feature vector $\boldsymbol{Z} = \texttt{concat}(\{\texttt{PLFS}\left(\boldsymbol{F}, S^o\right)\}) \in \mathbb{R}^{Kd}$. The updating process can be formulated as follows:

$$\begin{aligned} \Delta S &= \texttt{U}(\boldsymbol{Z}) = \boldsymbol{Z} \times W_U + b_U, \\ S &= S^o + \Delta S \end{aligned} \tag{7}$$

where $W_U \in \mathbb{R}^{Kd\times K\times 2}$ and $b_U \in \mathbb{R}^{K\times 2}$ are the updater's learnable parameters.

With the updated $S = \{(x,y)_k\}$, the local features of the next routing can be extracted $\{\boldsymbol{z}_{\texttt{next},k}\} = \texttt{PLFS}\left(\boldsymbol{F}_{\texttt{next}}, S\right) \in \mathbb{R}^{K\times d}$. The router estimates the coordinate offsets between each point to its finer counterparts. The routing process can be formulated as follows:

$$\begin{aligned} \{(\Delta x^o_{\texttt{next}}, \Delta y^o_{\texttt{next}})_{k,m}\} &= \texttt{R}(\boldsymbol{z}_{\texttt{next},k}) = \boldsymbol{z}_{\texttt{next},k} \times W_R + b_R, \\ \{(x^o_{\texttt{next}}, y^o_{\texttt{next}})_{k,m}\} &= \{(x,y)_k + (\Delta x^o_{\texttt{next}}, \Delta y^o_{\texttt{next}})_{k,m}\} \end{aligned} \tag{8}$$

where $W_R \in \mathbb{R}^{d\times M\times 2}$ and $b_R \in \mathbb{R}^{M\times 2}$ are the router's learnable parameters, and $m = 1, \ldots, M$.

The router is shared to spread each point onto M points. Thus, S finally becomes $S^o_{\texttt{next}}$ containing more points and finer semantics, which can further be updated and routed till the final positions of all landmarks are precisely localized.

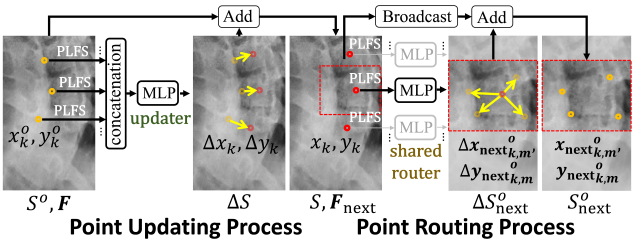

Figure 4: Details of point updating and routing. PLFS-extracted features are concatenated for first updating the point positions. The local features from new positions are then individually processed by a shared router to 'split' each point into multiple finer points.

Implementation and Training Details

The objective of RouterNet is getting close to the GT regression targets of different semantics, i.e., S_{rt}, S_{ctr}, S_{cnr}. The loss for each level of semantic is the total L1 distance between the updated and target points. Unlike the classic point regression which optimizes each regressor independently, PLFS allows the parameters of all updaters and routers as well as the multi-scale feature extractor to be optimized jointly. Therefore, we add the losses together for an end-to-end training. The overall loss function can be formulated as:

$$\mathcal{L}_{total} = |S_{rt} - \hat{S}_{rt}| + |S_{ctr} - \hat{S}_{ctr}| + |S_{cnr} - \hat{S}_{cnr}| \tag{9}$$

We use data augmentation (horizontal flipping, affine transformation, Gaussian blurring, Gamma transformation), Adam optimizer (lr=1e-3, halved quarterly), batch size 40, and 10 epochs.

Dataset and Evaluation Metrics

Dataset

This work includes two AP X-ray datasets, i.e., AASCE[1] (public) and our private dataset, and a public lateral X-ray dataset, i.e., NHANES II[2].

AASCE consists of 707 spinal AP X-ray images from the London Health Sciences Center in Canada. The images were

[1]https://aasce19.github.io/\#challenge-dataset

[2]https://wwwn.cdc.gov/nchs/nhanes/nhanes2/default.aspx

Method	$AASCE_{val}$			$AASCE_{test}$			Private		
	NMSE ↓	OR_{20} ↓	OR_{40} ↓	NMSE ↓	OR_{20} ↓	OR_{40} ↓	NMSE ↓	OR_{20} ↓	OR_{40} ↓
SCN	4.16E-3	45.14	22.35	5.32E-3	40.32	19.03	5.08E-3	34.34	15.74
SLSN	3.01E-3	49.60	23.22	3.40E-3	63.60	25.35	4.13E-3	48.88	22.66
HRNet	1.82E-3	36.35	15.40	2.65E-3	27.16	10.83	1.26E-3	17.95	7.61
Yi *et al.*	1.29E-3	29.33	11.26	1.66E-3	31.95	11.06	9.51E-4	18.94	6.19
H3R	1.25E-3	33.13	8.84	8.43E-4	24.56	5.87	9.52E-4	27.64	4.33
FARNet	1.15E-3	29.03	9.87	7.04E-4	14.99	4.73	8.20E-4	22.75	4.46
3000FPS	8.13E-4	12.26	2.16	1.43E-3	14.62	3.14	1.23E-3	13.88	3.68
Cascaded CNNs	4.88E-4	3.62	0.36	7.78E-4	6.68	0.96	3.38E-4	3.68	**0.04**
RouterNet (Ours)	**1.28E-4**	**1.72**	**0.08**	**2.00E-4**	**2.16**	**0.24**	**1.26E-4**	**1.17**	0.09

Table 1: Comparison results of vertebral landmark localization on three datasets, i.e., $AASCE_{val}$, $AASCE_{test}$ and Private. The first 6 methods are heatmap-based and the last 3 methods uses point-regression strategy. The best performance is marked in bold. See supplementary material for detailed results.

officially divided into 481, 128, and 98 for training, validation, and test, respectively. The training and validation images have the GT labels, each of which contains 4 corner points for each of 17 vertebrae, resulting in 68 landmarks in total. Since the official labels of 98 test images are unavailable, we invited two local experts to label them manually and the annotations are released with our codes for the research purpose.In the following, we denote the 128 validation and 98 test images as $AASCE_{val}$ and $AASCE_{test}$, respectively.

Our private dataset contains 36 AP X-ray images collected from a local hospital. Each image was scanned from a scoliosis patient using the Canon X-ray system. The same two local experts were invited to annotate the 68 landmarks of each image, which are consistent with those in the $AASCE$.

NHANES II contains 214 cervical annotated images collected from 1976 to 1980 conducted by the NCHS for the Second National Health and Nutrition Examination Survey. The images were annotated with 4 vertebrae (C2-C5, 16 corner landmarks). We randomly divided the images into 171 (80%) and 43 (20%) for training and test, respectively.

Evaluation Metrics

For the task of landmark localization, we employ the consistent metrics with the previous works, i.e., Normalized Mean Squared Error (NMSE), which is calculated as:

$$NMSE = \frac{1}{K}\sum_{k=1}^{K}(\frac{x_k - \hat{x}_k}{W})^2 + (\frac{y_k - \hat{y}_k}{H})^2 \quad (10)$$

where K is the number of landmarks, i.e., 68 for AASCE and our Private, and 16 for NHANES II, (x_k, y_k) and $(\hat{x}_k, \hat{y}_k)$ are the predicted and GT absolute point coordinates, and W and H are the width and height of image.

We also report Outlier Ratio outside the distance of r pixels (OR_r) to evaluate the missing/false points, calculated as:

$$OR_r = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}((x_k - \hat{x}_k)^2 + (y_k - \hat{y}_k)^2 > r^2) \quad (11)$$

where $\mathbb{I}(.)$ is 1 if the inner equation is true and 0 if false, and r defines an error radius.

Besides, AASCE was initialized for a challenge of downstream task called Cobb angle estimation and provided the GT of three angles, i.e., the proximal thoracic (PT), main thoracic (MT), and thoracolumbar (TL) angles. Therefore, we convert the predicted landmarks to the Cobb angle biomarker using the AASCE official tool[3], and employ the five challenge-adopted metrics for evaluation, i.e., Manhattan Distance (MD), Euclidean Distance (ED), Chebyshev Distance (CD), Circular Mean Absolute Error (CMAE), and Symmetric Mean Absolute Percentage Error (SMAPE). The calculation of these metrics can refer to the challenge paper (Wang et al. 2021).

Experimental Results and Discussions

AP-view Vertebral Landmark Localization

We choose 8 recent state-of-the-arts for comparison of vertebral landmark localization, i.e., Yi *et al.* (Yi et al. 2020), FARNet (Ao and Wu 2023), SLSN (Zhang et al. 2021), Cascaded CNNs (Wang et al. 2024), SCN (Payer et al. 2019), H3R (Yu and Tao 2021), HRNet (Sun et al. 2019), and 3000FPS (Ren et al. 2016). Among them, Cascaded CNNs (Wang et al. 2024) and 3000FPS (Ren et al. 2016) belong to point-regression and the rest ones are heatmap-based. All comparison methods have released source codes. We train the comparison methods and ours using the training data (481 images) of AASCE, and perform the evaluation on $AASCE_{val}$, $AASCE_{test}$, and Private, respectively.

The comparison results are listed in Table 1. Point-regression methods consistently outperform heatmap-based approaches due to inherent shape constraints that eliminate missing/false detections critical for clinical applications. RouterNet achieves significant improvements: **73.8%** NMSE reduction on $AASCE_{val}$ compared to Cascaded CNNs (Wang et al. 2024) and **71.6%** on $AASCE_{test}$ compared to FARNet (Ao and Wu 2023), with near-zero missing/false detections.

[3]http://spineweb.digitalimaginggroup.ca/

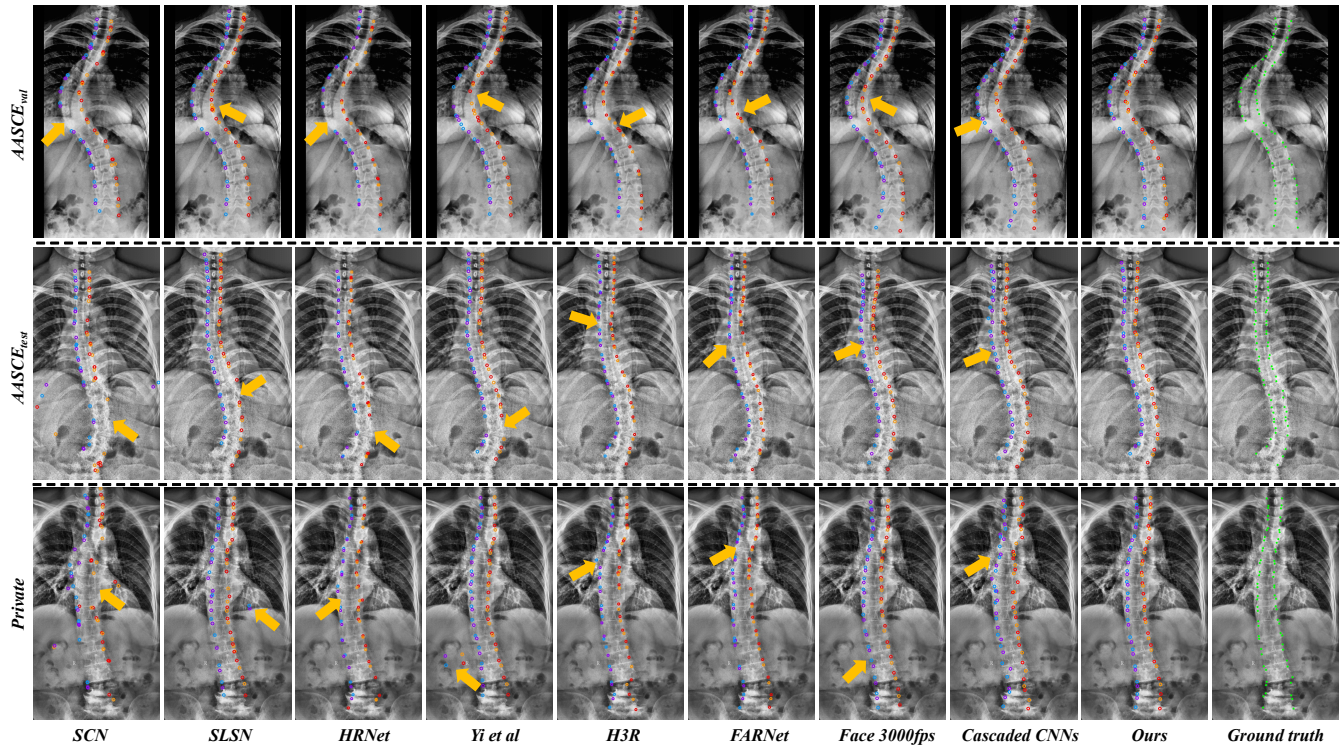

Figure 5: Visualization results on $AASCE_{val}$, $AASCE_{test}$ and our private dataset. Different colors indicate the four corner points for each vertebra.

To further verify the cross-dataset generalization performance of the compared methods, we directly apply the best model evaluated on $ASSCE_{val}$ to predict our private dataset. On Private dataset, RouterNet still achieves **62.7%** NMSE improvement over Cascaded CNNs, demonstrating consistent performance across datasets.

Fig. 5 gives visual results of different methods on $AASCE_{val}$, $AASCE_{test}$ and our private dataset. It can be observed that the heatmap-based methods predict many missing/false points. In comparison, this error rarely occurs in RouterNet's results. To further reveal this, we plot curves of detection rate by varying normalized error radius(see supplementary material). The curves show that RouterNet achieves near-perfect detection rates with minimal missing/-false predictions compared to heatmap-based methods.

Lateral-view Cervical Landmark Localization

To verify the applicability, we compare our method with others on the cervical lateral X-ray images from NHANES II, and the comparison results are shown in Table 2. We can observe that the performance on this dataset is better than that on AP X-ray images for all methods. This is due to the fact that the lateral cervical X-ray images contain fewer vertebra-like structures from the heart and lung, and the advantage of using heatmaps emerges, that is, HRNet (Sun et al. 2019) becomes the second-best. Nevertheless, RouterNet still outperforms others in terms of all metrics, indicating its good applicability to other body parts.

Method	NMSE ↓	OR_{15} ↓	OR_{30} ↓
SCN	1.55E-4	3.20	2.91
SLSN	9.47E-5	0.73	0.73
HRNet	2.87E-5	0.29	0.15
Yi *et al.*	1.12E-4	1.45	1.16
H3R	2.53E-4	6.40	3.78
FARNet	8.55E-5	1.60	1.31
3000FPS	2.93E-5	0.87	0.00
Cascaded CNNs	1.03E-4	0.73	0.00
RouterNet (Ours)	**1.74E-5**	**0.00**	**0.00**

Table 2: Comparison results of cervical landmark localization on NHANES II. The best performance is marked in bold.

Ablation Studies

The Effectiveness of Hierarchical Routing Table 3 validates hierarchical routing effectiveness through three variants: direct 68-landmark regression without routing, routing once (centers→corners), and routing twice (root→centers→corners). Progressive routing achieves 61.7% and 60.6% NMSE improvements respectively, confirming the divide-and-conquer strategy's effectiveness.

The Selection of Root Point We also analyze the impact of different selections of root point. We exhaustively train

Method	NMSE ↓	OR_{20} ↓	OR_{40} ↓
w/o Routing	8.48E-4	23.98	2.83
Routing once	3.25E-4	4.39	0.13
Routing twice	1.28E-4	1.72	0.08

Table 3: The comparison results on $AASCE_{val}$ of RouterNet ('Routing twice') and its two variants with different routing times.

Method	CMAE ↓	ED ↓	MD ↓	CD ↓	SMAPE ↓
Wang *et al.*	-	-	-	-	23.43
Horng *et al.*	-	-	-	-	16.48
PFA	6.69	-	-	-	12.97
CGN	4.77	-	-	-	10.25
Guo *et al.*	-	-	-	-	8.62
Kpt-Transformer	-	-	-	-	8.40
W-Transformer	-	-	-	-	8.26
VF	3.51	-	-	-	7.84
Seg4Reg	3.96	-	-	-	7.64
Seg4Reg+	3.73	-	-	-	7.32
SCN	22.80	44.08	68.42	34.74	31.62
SLSN	4.04	8.19	12.14	6.86	9.08
HRNet	12.23	24.38	36.71	19.53	21.77
Yi *et al.*	3.55	7.13	10.63	5.90	8.21
H3R	7.20	14.60	21.60	12.12	15.34
FARNet	8.06	16.38	24.17	13.43	16.15
3000FPS	7.61	15.23	22.84	12.64	17.01
Cascaded CNNs	6.51	12.81	19.54	10.37	14.60
RouterNet (Ours)	**2.69**	**5.30**	**8.07**	**4.27**	**6.24**

Table 4: Comparison results of scoliosis assessment on $AASCE_{val}$.The best performance is marked in bold.

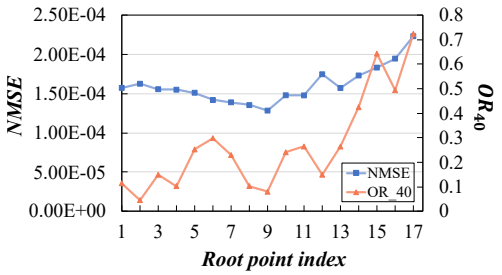

Figure 6: The effect of selecting different vertebra as root point on $AASCE_{val}$.

17 variants of RouterNet by using the first to last vertebra's center point as root, and the comparison results are shown in Fig. 6. Analysis of root point selection shows the 9th vertebra center achieves optimal performance, enabling potential semi-automatic clinical deployment.

Application on Scoliosis Assessment

We evaluate our method in a downstream task called scoliosis assessment on $AASCE_{val}$ using the officially released GT Cobb angles.

Method	CMAE ↓	ED ↓	MD ↓	CD ↓	SMAPE ↓
XMU	4.91	11.23	14.74	10.17	22.18
iFLYTEK	5.48	12.14	16.45	10.74	22.17
Seg4Reg	4.85	11.17	14.55	10.16	21.71
SCN	29.79	58.34	89.35	46.11	48.26
SLSN	3.59	7.39	10.78	6.21	11.51
HRNet	17.64	34.72	52.99	27.64	36.59
Yi *et al.*	3.42	6.84	10.25	5.61	11.44
H3R	8.95	17.43	26.85	14.04	23.48
FARNET	8.61	17.37	25.77	14.20	21.63
3000FPS	7.58	14.90	22.75	12.01	23.68
Cascaded CNNs	7.46	14.83	22.38	12.24	21.30
RouterNet (Ours)	**2.80**	**5.63**	**8.40**	**4.64**	**9.45**

Table 5: Comparison results of scoliosis assessment on $AASCE_{test}$. The best performance is marked in bold.

For the localization methods, the Cobb angles are calculated using the official tool as mentioned in the evaluation metrics section. Besides, we also include 10 recent state-of-the-arts focusing on scoliosis assessment. The results of (Huo et al. 2021; Guo et al. 2021a,b; Yao et al. 2022; Lin et al. 2019, 2021) are reported in their original papers, and the results of (Wang et al. 2019; Horng et al. 2019; Wang, Wang, and Liu 2019; Kim et al. 2020) are borrowed from (Yao et al. 2022; Lin et al. 2021).

The comparison results are listed in Table 4. RouterNet achieves the best scoliosis assessment performance, outperforming VF (Kim et al. 2020) by 23.4% (CMAE) and Seg4Reg+ (Lin et al. 2021) by 14.8% (SMAPE).

On challenge test data $AASCE_{test}$, we include the results of the top three participants(Lin et al. 2019; Chen et al. 2019; Wang, Huang, and Wang 2019). The comparison results are listed in Table 5. Our method achieves SMAPE less than half of the champion's (Lin et al. 2019).

Conclusion

We propose RouterNet, a hierarchical point routing network that decomposes vertebral landmark localization into progressive sub-tasks. By coupling multi-scale feature learning and point-guided local feature selection (PLFS) with coarse-to-fine point routing, RouterNet achieves superior performance while eliminating missing/false detections. Extensive experiments on three datasets demonstrate significant improvements in both landmark localization and scoliosis assessment, with potential for clinical deployment through semi-automatic operation.

Although RouterNet is fully automatic in this work by use of the initial root pre-calculated based on all training samples, it can be naturally extended to be semi-automatic and more controllable in the clinical practice by allowing the doctor to give an initial point manually. Our further work will focus on vertebral landmark localization on X-ray images with arbitrary FOV.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No.62202189), research grants from Wuhan United Imaging Healthcare Surgical Technology Co., Ltd.

References

Ao, Y.; and Wu, H. 2023. Feature Aggregation and Refinement Network for 2D Anatomical Landmark Detection. *Journal of Digital Imaging*, 36(2): 547–561.

Cao, X.; Wei, Y.; Wen, F.; and Sun, J. 2014. Face alignment by explicit shape regression. *International journal of computer vision*, 107(2): 177–190.

Chen, K.; Peng, C.; Li, Y.; Cheng, D.; and Wei, S. 2019. Accurate automated keypoint detections for spinal curvature estimation. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, 63–68. Springer.

Guo, Y.; Li, Y.; He, W.; and Song, H. 2021a. Heterogeneous Consistency Loss for Cobb Angle Estimation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2588–2591. IEEE.

Guo, Y.; Li, Y.; Zhou, X.; and He, W. 2021b. A keypoint transformer to discover spine structure for cobb angle estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Horng, M.-H.; Kuok, C.-P.; Fu, M.-J.; Lin, C.-J.; and Sun, Y.-N. 2019. Cobb angle measurement of spine from X-ray images using convolutional neural network. *Computational and mathematical methods in medicine*, 2019.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huo, L.; Cai, B.; Liang, P.; Sun, Z.; Xiong, C.; Niu, C.; Song, B.; and Cheng, E. 2021. Joint Spinal Centerline Extraction and Curvature Estimation with Row-Wise Classification and Curve Graph Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 377–386. Springer.

Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, 784–799.

Kim, K. C.; Yun, H. S.; Kim, S.; and Seo, J. K. 2020. Automation of spine curve assessment in frontal radiographs using deep learning of vertebral-tilt vector. *IEEE Access*, 8: 84618–84630.

Lin, Y.; Liu, L.; Ma, K.; and Zheng, Y. 2021. Seg4Reg+: Consistency Learning Between Spine Segmentation and Cobb Angle Regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 490–499. Springer.

Lin, Y.; Zhou, H.-Y.; Ma, K.; Yang, X.; and Zheng, Y. 2019. Seg4Reg networks for automated spinal curvature estimation. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, 69–74. Springer.

Matthews, I.; and Baker, S. 2004. Active appearance models revisited. *International journal of computer vision*, 60(2): 135–164.

Payer, C.; Štern, D.; Bischof, H.; and Urschler, M. 2019. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical image analysis*, 54: 207–219.

Ren, S.; Cao, X.; Wei, Y.; and Sun, J. 2016. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3): 1233–1245.

Sun, H.; Zhen, X.; Bailey, C.; Rasoulinejad, P.; Yin, Y.; and Li, S. 2017. Direct estimation of spinal cobb angles by structured multi-output regression. In *International conference on information processing in medical imaging*, 529–540. Springer.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.

Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; and Le, Q. V. 2019. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2820–2828.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; and Bregler, C. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 648–656.

Wang, J.; Jin, Y.; Cai, S.; Xu, H.; Heng, P.-A.; Qin, J.; and Wang, L. 2022. Real-time landmark detection for precise endoscopic submucosal dissection via shape-aware relation network. *Medical Image Analysis*, 75: 102291.

Wang, J.; Wang, L.; and Liu, C. 2019. A multi-task learning method for direct estimation of spinal curvature. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, 113–118. Springer.

Wang, L.; Xie, C.; Lin, Y.; Zhou, H.-Y.; Chen, K.; Cheng, D.; Dubost, F.; Collery, B.; Khanal, B.; Khanal, B.; et al. 2021. Evaluation and comparison of accurate automated spinal curvature estimation algorithms with spinal anterior-posterior X-Ray images: The AASCE2019 challenge. *Medical Image Analysis*, 72: 102115.

Wang, L.; Xu, Q.; Leung, S.; Chung, J.; Chen, B.; and Li, S. 2019. Accurate automated Cobb angles estimation using multi-view extrapolation net. *Medical Image Analysis*, 58: 101542.

Wang, S.; Huang, S.; and Wang, L. 2019. Spinal curve guide network (SCG-Net) for accurate automated spinal curvature estimation. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, 107–112. Springer.

Wang, Z.; Lv, J.; Yang, Y.; Lin, Y.; Li, Q.; Li, X.; and Yang, X. 2024. Accurate scoliosis vertebral landmark localization on X-ray images via shape-constrained multi-stage cascaded CNNs. *Fundamental Research*, 4(6).

Wu, H.; Bailey, C.; Rasoulinejad, P.; and Li, S. 2017. Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, 127–135. Springer.

Yao, Y.; Yu, W.; Gao, Y.; Dong, J.; Xiao, Q.; Huang, B.; and Shi, Z. 2022. W-Transformer: Accurate Cobb angles estimation by using a transformer-based hybrid structure. *Medical Physics*.

Yi, J.; Wu, P.; Huang, Q.; Qu, H.; and Metaxas, D. N. 2020. Vertebra-focused landmark detection for scoliosis assessment. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 736–740. IEEE.

Yu, B.; and Tao, D. 2021. Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8276–8289.

Zhang, C.; Wang, J.; He, J.; Gao, P.; and Xie, G. 2021. Automated vertebral landmarks and spinal curvature estimation using non-directional part affinity fields. *Neurocomputing*, 438: 280–289.