WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n1.12
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Cephalometric Landmark Localization Model Based on Polarized Self-Attention Mechanism

Shuaichao Feng [1], Xinpeng Miao [1], Shukui Ma [1], Fei Ma [2], Guangping Zhuo [1, *]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China

[2] Department of Computer Science and Technology, Taiyuan University, Taiyuan 030032, China

ABSTRACT

Precise localization of cephalometric landmarks is crucial in the fields of orthodontics and craniofacial surgery. Traditional manual cephalometric analysis and computer-aided cephalometric analysis have significant drawbacks, including large errors, low accuracy, and being time-consuming. To achieve efficient and accurate localization of cephalometric landmarks, this study proposes a detection algorithm, CenterNet-PSA, which integrates the Polarized Self-Attention Mechanism. The algorithm first uses a pre-trained DLA-34 as the feature extraction network to extract features, and then incorporates the polarized self-attention mechanism into the DLA-34 feature extraction network to weight the spatial and channel information of the image, thereby improving the accuracy of landmark detection. Finally, the model achieves a mean radial error (MRE) of 1.07mm and a success detection rate (SDR) of 88.14% within a 2mm error range on the ISBI 2015 Grand Challenge cephalometric X-ray test dataset. Compared to other detection methods, CenterNet-PSA can achieve efficient and accurate localization of cephalometric landmarks, meeting the needs of clinical medicine.

KEYWORDS

Orthodontics; Cephalometric Landmark; DLA-34; Polarized Self-Attention Mechanism

1. INTRODUCTION

Oral health is a crucial component of overall health, and the World Health Organization (WHO) includes it as one of the top ten health standards. It not only reflects the physical and mental health status of the population in a country or region but also indicates the level of social civilization. The government pays significant attention to oral health and has introduced a series of policies and laws to promote the continuous development of oral health care.

In recent years, with the increasing efforts to promote oral health awareness, the public's awareness of oral health in the country has gradually improved. However, there are still challenges in clinical dentistry: initial diagnostic data is complex to calculate, manual measurements lack precision, and the process is time-consuming. Initial data collection and measurement include X-rays (with at least 30 measurement items after refinement), plaster models, facial and intraoral images, all of which require manual measurements by doctors. The inherent variability in the quality of head X-ray imaging and the complex differences in individual anatomical structures pose significant challenges for reliable landmark annotation. Even for experienced orthodontists, manually identifying these landmarks remains a labor-intensive and time-consuming task. Therefore, automatic and accurate localization of cephalometric landmarks is of great importance in clinical practice, especially for orthodontic diagnosis and treatment planning.

In the early stages of cephalometric landmark detection research, researchers predominantly relied on classical image processing techniques. For instance, the Grau team [1] was among the first to use template matching techniques to determine key point locations in cephalometric measurements. However, as image complexity increased, the stability and effectiveness of template matching began to show limitations. To address this issue, researchers like Keustermans et al. [2] shifted to automatic detection techniques based on local appearance features for determining these key points. On the other hand, scholars such as Ibragimov et al. [3] adopted game theory and morphological models to extract features from X-ray images. Despite these advancements, these techniques often required significant human labor and had certain limitations in performance. Subsequently, studies [4-6] introduced machine learning algorithms like support vector machines and random forests for landmark localization. By integrating local image details with global information such as organ size and posture, these methods effectively improved prediction accuracy.

With the rapid advancement of deep learning technologies, their application in medical image processing, such as anatomical structure key point localization, has become increasingly widespread. In 2017, Lee et al. [7] were the first to propose a key point localization method using convolutional neural networks (CNNs) for coordinate regression. They converted the x and y coordinates of key points into a 38-dimensional 1D vector and established a multivariate regression system to predict the coordinate values. In 2019, Zhong et al. [8] developed a model combining deep encoder-decoder architecture, which integrated both global key point locations and local high-resolution features. They adopted a two-stage U-shaped network structure and performed key point localization through multi-channel heatmap regression. Additionally, an attention mechanism was incorporated into the global phase heatmap to guide local phase inference and enable high-resolution heatmap regression. Qian et al. [9], based on the unique attributes of cephalometric images and the distribution features of key points, developed a fast detection model, CephaNet, based on the Region Convolutional Neural Network (R-CNN) [10]. In the CephaNet model, they designed a multi-task loss function aimed at reducing intra-class differences and enhanced the detection ability of small key points through multi-scale training strategies. Dai et al. [11] proposed a cephalometric key point localization technique based on Generative Adversarial Networks (GAN [12]), which generated distance maps of key points using GANs and determined the key point coordinates through a regression voting mechanism. When dealing with cephalometric landmark localization regression tasks, the commonly used encoder-decoder structures often lead to performance degradation due to reduced spatial dimensions and increased channel dimensions. Most existing methods also utilize relatively shallow networks for feature extraction, failing to provide high-resolution feature maps, which results in quantization errors between predicted and actual values.

To address the limitations of previous methods, this study developed an upgraded single-stage detection algorithm based on CenterNet-PSA. The following key improvements were made to the CenterNet network: (1) A series of data augmentation strategies were implemented, including geometric transformations and color space adjustments, such as rotation, histogram equalization, and image sharpening, to enhance the model's overall performance and generalization ability; (2) The Polarized Self-Attention (PSA [13]) module was integrated into the feature extraction network, enhancing the model's understanding of features and improving its global perception of both channel and spatial positions, as well as its ability to capture local detail information; (3) Dropout technology was introduced to effectively reduce the risk of overfitting, thereby improving the model's performance on the test dataset. The model demonstrated excellent performance on a public X-ray cephalometric dataset and shows significant potential for clinical application.

2. METHOD

2.1. Model Overview

The overall framework of the method in this paper is shown in Figure 1. CenterNet is a single-stage object detection algorithm that abandons traditional bounding boxes and anchor mechanisms, instead modeling the object as a single point, i.e., the center point of the bounding box. By detecting key points to predict the center of the object, and then combining size and offset information to reconstruct the bounding box, this method simplifies the detection process, eliminates post-processing steps, and improves both efficiency and accuracy. The network primarily consists of a feature extraction network and a center point prediction network. The feature extraction network typically employs pre-trained convolutional neural networks (such as ResNet [14], Hourglass [15], or DLA-34 [16]) to extract image features. After continuous convolution and pooling operations on the feature map, the center point prediction network performs further convolution and upsampling operations to restore the feature map's dimensions to match the original image and generates a heatmap. In this heatmap, the point with the highest value represents the center location of the target object. In this way, the network is able to predict the center coordinates, bounding box size, and offset parameters of each target object.

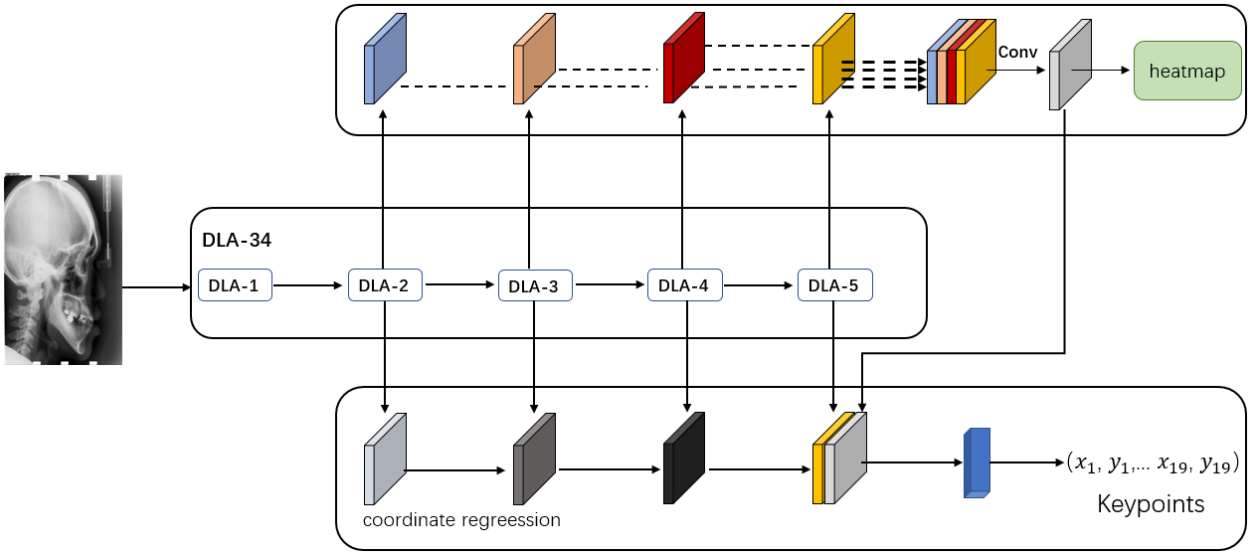

Figure 1. CenterNet network structure

This paper primarily uses the Deep Layer Aggregation Network (DLA-34) as the backbone for feature extraction. DLA-34 emphasizes feature aggregation at various levels of the network, allowing for better integration of low-level detailed information and high-level semantic features. This structure enables DLA-34 to capture richer, multi-scale feature information, effectively enhancing the model's perception capability and robustness, while maintaining relatively low computational complexity. The network structure is shown in Figure 2.

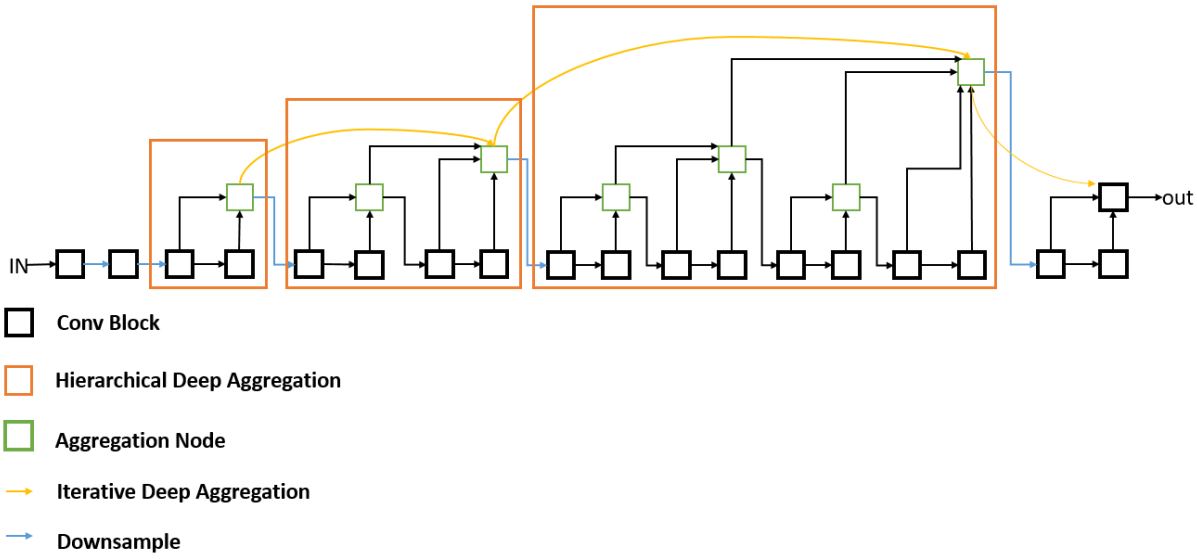

Figure 2. DLA-34 structure diagram

First, the input image is processed through the pre-trained DLA-34 feature extraction network, which progressively extracts feature maps at different depths. These feature maps contain various levels of image information, ranging from low-level details such as edges and textures to high-level semantic information. DLA-34 then integrates these different layers of features through iterative depth aggregation and hierarchical depth aggregation to combine both semantic and spatial information. Hierarchical depth aggregation merges deep and shallow features within the same scale, gradually constructing deeper and richer feature representations. This enhances the network's ability to express and capture features within the same scale, thereby improving the model's capacity to capture fine-grained information. The core of iterative depth aggregation lies in mapping features from different scales to the same size. By fusing cross-scale information, it leverages high-level semantic data while preserving low-level detail features, thus strengthening the model’s understanding of both contextual information and local details.

2.2. Model Improvement

In this paper, the feature extraction network in the CenterNet model is improved, with a focus on modifying the structure of the DLA-34 network. Additionally, the ReLU activation function in the network is replaced with LeakyReLU, which effectively alleviates the Dead ReLU problem. The DLA-34 network is primarily composed of basic convolutional layers, aggregation layers, and depth aggregation layers. In the basic convolutional layer of DLA-34, after the initial convolution, batch normalization (BN), and LeakyReLU activation, a Polarized Self-Attention (PSA) module is introduced. Subsequently, the output processed by the PSA module undergoes a second convolution and BN, and is added to the downsampled input. After LeakyReLU activation, the final output is obtained. The structure is shown in Figure 3.

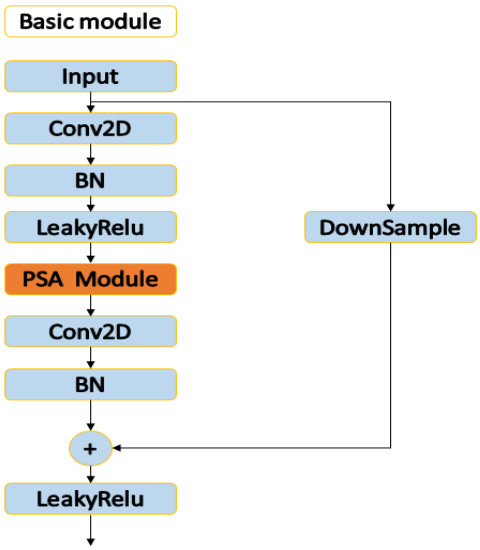

Figure 3. Improvement of DLA-34 basic module

In this study, a similar processing flow is applied to the aggregation layers of DLA-34. Initially, the aggregation layer undergoes convolution, batch normalization (BN), and LeakyReLU activation. The resulting output is then fed into the PSA module to capture the aggregated spatial and channel-wise features. Next, the output from the PSA module is added to the initial input and processed through LeakyReLU activation to produce the final output. The specific architecture is detailed in Figure 4.

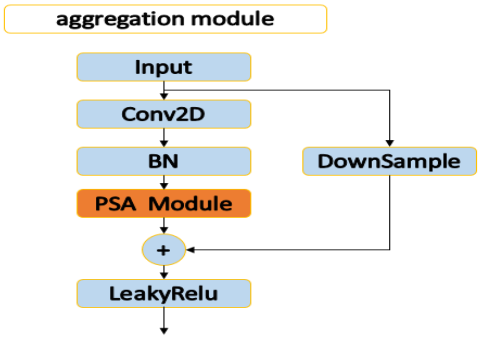

Figure 4. Improvement of DLA-34 aggregation module

2.3. Polarized Self-Attention Mechanism

The Polarized Self-Attention Mechanism (PSA) is an improved self-attention mechanism designed to enhance computational efficiency and improve the model's performance in processing image features. The structure is shown in Figure 5. The self-attention mechanism operates on the input data to either amplify or suppress specific features, similar to how optical lenses filter light. In photography, scattered light often causes glare or reflections in the horizontal direction. Allowing only light that is perpendicular to the horizontal direction to pass through may enhance the image contrast. The filtered light, with reduced overall brightness, often exhibits a narrower dynamic range. Therefore, further enhancement techniques, such as High Dynamic Range (HDR) imaging, may be required to restore scene details. This analogy is applied in the PSA mechanism, where the model adjusts feature importance based on spatial and channel information. The PSA mechanism helps the model focus on the most relevant features while suppressing irrelevant ones, ultimately improving the model's ability to capture both fine-grained and high-level information for tasks such as landmark detection.

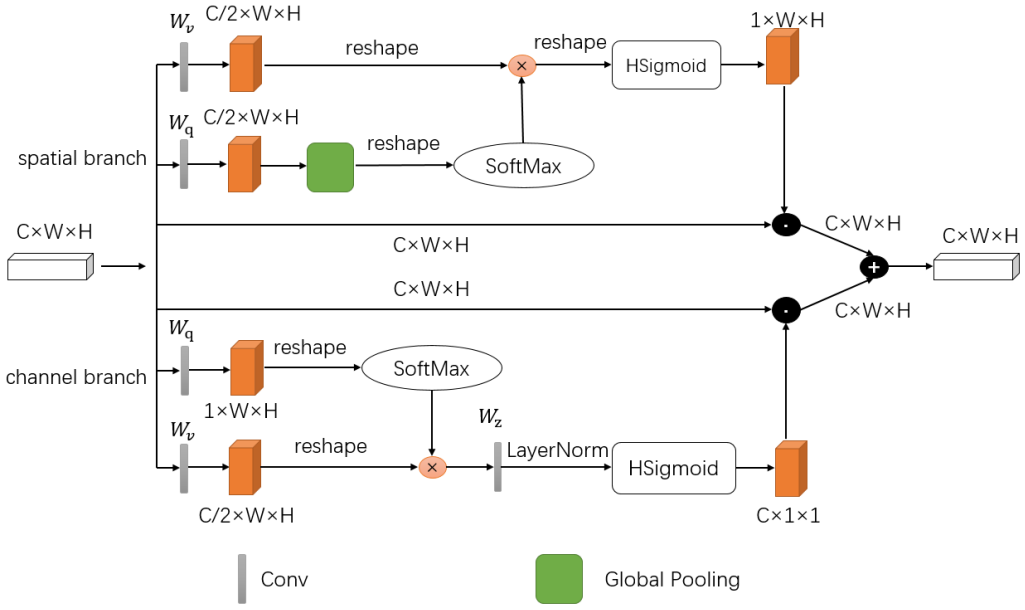

Figure 5. Polarized Self-Attention module

The Polarized Self-Attention Mechanism (PSA) has two core components:

Polarized Filtering: In most pixel-level regression algorithms, to enhance robustness and reduce computational complexity, low-resolution feature maps are often generated. This approach may lead to the loss of complex edge details of objects. The polarized self-attention mechanism, through its unique polarized filtering process, achieves complete compression of features along one dimension while preserving high-resolution characteristics along the perpendicular dimension. For example, when the channel dimension is compressed, the spatial dimension retains its high resolution; conversely, when the spatial dimension is compressed, the channel dimension maintains high resolution. This strategy significantly improves the accuracy of keypoint localization.

HDR Mechanism: Within the self-attention module, a softmax normalization operation is applied to the smallest scale feature tensor to expand the attention range and enhance the expression of information. Following this, an HSigmoid function is used to complete the feature mapping transformation. This mechanism enables the model to focus on the most critical features while enhancing the overall performance of the attention mechanism.

The core idea of the Polarized Self-Attention Mechanism (PSA) is to divide the attention calculation into two directions: horizontal (lateral) and vertical (perpendicular). This approach differs from standard global self-attention, which computes the mutual interactions between every position in the two-dimensional space. By decomposing the attention into these two directions, the model can effectively capture spatial contextual information while reducing computational cost. Additionally, in this study, the Sigmoid function in the PSA mechanism is replaced with the HSigmoid function. This change improves computational speed, reduces complexity, and effectively alleviates the gradient vanishing problem.

In the channel branch, the determination of weights follows a specific computational rule, which is expressed as follows:

$$A^{ch}(\mathrm{X}) = F_{SG}\left[W_{z|\theta_1}\left(\left(\sigma_1\left(W_v(X)\right) \times F_{SM}\left(\sigma_2\left(W_q(X)\right)\right)\right)\right)\right] \tag{1}$$

In the above formula, W_q, W_v and W_z represent 1×1 convolution layers, σ_1 and σ_2 are two different dimensionality reduction operations, F_{SG} refers to the Hsigmoid function, F_{SM} refers to the softmax function. The × symbol denotes matrix dot product operation. In the channel branch

processing, the W_v convolution layer reduces the number of channels by half, then the 2D feature map is transformed into 1D. After that, a dot product operation is performed with the compressed spatial features, and the number of channels is restored through the W_z convolution layer. Finally, the Sigmoid function is applied to normalize the result, redistributing the weights back to the different channels of the original features. At the same time, the compressed spatial information undergoes information enhancement through the F_{SM} function.

The following is the specific calculation method for the spatial branch weights:

$$A^{sp}(\mathrm{X}) = F_{SG}\left[\sigma_3(F_{SM}(\sigma_1\left(F_{GP}\left(W_q(X)\right)\right)) \times \sigma_2(W_v(X)))\right] \tag{2}$$

W_q and W_v are 1×1 convolution kernels with one−dimensional size, σ_1 and σ_2 are responsible for performing dimensionality reduction, σ_3 performs dimensionality expansion, F_{SM} is the softmax function, $F_{GP}(\cdot)$ represents the global pooling operation.

The differences between the spatial branch and the channel branch are mainly reflected in the following two aspects:

In the spatial branch, all the channel information across spatial dimensions is compressed, and then the convolution, global pooling, dimensionality reduction, and softmax functions are applied to complete the regression task.

After the dot product, no additional convolution processing is added. Instead, dimensionality recovery is directly performed using σ_3.

Finally, the processing results of these two branches are fused together to produce the output of the polarized self-attention mechanism. In the parallel structure, the outputs of the two branches are combined in the following form:

The outputs of the above two branches are composed in a parallel layout as follows:

$$\mathrm{PSA}_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \odot^{ch} X + A^{sp}(X) \odot^{sp} X \tag{3}$$

It can be seen that PSA maintains high-resolution characteristics in both the spatial and channel dimensions, and applies nonlinear activation to the tensor of the bottleneck layer using the softmax function. By adjusting the weights, that is, by assigning higher weights to key features, the representation of significant features in the image can be enhanced, leading to more precise localization of the cephalometric landmarks.

2.4. Loss Function

The loss function is a tool to measure the deviation between predicted and actual values, guiding the adjustment of model parameters to improve model performance. For the CenterNet-PSA architecture, this study adopts a weighted loss function as the overall loss measure during the training process. The formula for the weighted loss function is as follows:

$$L_{tot} = L_{hk} + \lambda_{wh} + \lambda_O L_O \tag{4}$$

Where L_{hk}, λ_{wh} and L_O represent the heatmap loss, width-height prediction regression loss (wh loss), and object center offset loss (reg loss), respectively. Specifically, the formula for the heatmap loss is as follows:

$$L_{hk} = -\frac{1}{N}\begin{cases} \left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right) & if\ Y_{xyc}=1 \\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}\right)^{\alpha}\log\left(1-\hat{Y}_{xyc}\right) & othersize \end{cases} \tag{5}$$

N represents the total number of key points, Y_{xyc} denotes the actual coordinates of the center point, and $\hat{Y}_{xyc}$ represents the predicted coordinates of the center point.

The loss function for the object center offset can be defined as follows:

$$L_O = -\frac{1}{N}\sum_P \left|\hat{B}_{\tilde{P}} - \left(\frac{P}{4} - \tilde{P}\right)\right| \tag{6}$$

p represents the center position of the target region, 4 denotes the downsampling factor, $\tilde{P}$ represents the predicted target center position, $\hat{B}_{\tilde{P}}$ denotes the predicted center offset, and $\frac{P}{4}$ refers to the actual coordinates of the target after scaling.

The loss function for the target width-height error is defined as follows:

$$L_{wh} = \frac{\sum_{k=1}^{N}\left|\hat{s}_{pk} - S_k\right|}{N} \tag{7}$$

Where $\hat{S}_{pk} \in R^{\frac{W}{R}\times\frac{H}{R}\times 2}$ represents the predicted width and height loss from the network, and s_k denotes the ground truth values.

3. EXPERIMENTAL DESIGN

3.1. Dataset

In this study, the head lateral X-ray images provided by the ISBI 2015 Challenge were selected as the experimental dataset. The dataset contains 400 head lateral X-ray images. Following the competition rules, 150 images were designated as the training set, while 100 and 150 images were used as Test Set 1 and Test Set 2, respectively. All images were annotated with 19 key points by two experienced physicians, and the average of the annotations from the two physicians was used as the ground truth for training and testing. The image size is unified to 1935×2400 pixels, with each pixel representing a real-world size of 0.1 millimeters. During the experimental phase, to improve model convergence efficiency, image normalization was applied. Specifically, the RGB channels of the input images were normalized using the following mean and standard deviation values: mean = [0.408, 0.448, 0.470] and standard deviation = [0.289, 0.274, 0.278]. This normalization process helps make the data distribution more uniform, thereby improving the training efficiency of the model.

3.2. Experimental Environment

The experiment in this paper is based on the Linux operating system, with hardware primarily consisting of 8 GeForce RTX 3080 GPUs. The software environment used for the experiment includes Python 3.8 and the Pytorch 1.11.0 framework. The backbone network is the modified Deep Layer Aggregation network (DLA-34). The key model parameters are as follows: the initial learning rate is set to 1.5e-4, decaying to 0.1 times the original value at the 200th and 400th epochs; the optimizer used is Adam; the dropout rate is set to 0.3; the total number of training epochs is 1200, with a batch size of 8.

3.3. Data Preprocessing

To avoid overfitting due to insufficient training data, we performed data augmentation on the training set, including geometric transformations, color space adjustments, and the addition of random noise. First, the X-ray images were scaled proportionally to a width and height of 1024 pixels, with zero-padding added to the image borders to maintain the original aspect ratio. Next, a random rotation operation was applied to the images with angles ranging from -30° to +30°. Lastly, the brightness of the images was randomly adjusted. For X-ray images with low contrast, histogram equalization was applied to enhance the contrast, along with the addition of random noise. These preprocessing operations improved the quality and diversity of the dataset, enabling the model to better learn effective features and enhance its accuracy, generalization, and robustness.

3.4. Evaluation Metrics

In the task of cranial measurement localization, the evaluation metrics for model performance mainly include the Mean Radial Error (MRE) and the Success Detection Rate (SDR) within thresholds of 2mm, 2.5mm, 3mm, and 4mm. The MRE refers to the Euclidean spatial distance between the predicted key points and the actual key points. On the other hand, SDR measures the proportion of predicted key points that are within a specific threshold. For a dataset with N images and M key points, the calculation of these two metrics can be expressed by equations (8) and (9) respectively:

$$\mathrm{MRE} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left\|x_{n,m} - \hat{x}_{n,m}\right\|_2 \tag{8}$$

$$\mathrm{SDR} = \frac{\left|\left\{(n,m)\left\|x_{n,m}-\hat{x}_{n,m}\right\|_2 \le r\right\}\right|}{NM} \times 100\% \tag{9}$$

In this formula, $x_{n,m}$ refers to the actual coordinates, $\hat{x}_{n,m}$ refers to the predicted coordinates, and r symbolizes the permissible error threshold, which can be 2mm, 2.5mm, 3mm, or 4mm. Experimental validation shows that a lower MRE value indicates higher localization accuracy of the model, while a higher SDR value reflects better model performance.

3.5. Experimental Comparison

This study follows the ISBI 2015 competition guidelines and uses the dataset provided by the competition for model training and validation. Additionally, a comparative study was conducted by evaluating the performance of several state-of-the-art algorithms on the same dataset. The specific experimental results are shown in Table 1.

Table 1. Comparison of CenterNet-PSA results with other methods

Method	Test Dataset				
	MRE/mm	Success Detection Rate/%			
		Localization Error 2.0mm	Localization Error 2.5mm	Localization Error 3.0mm	Localization Error 4.0mm
IBRAGIMOV	1.84	71.70	77.40	81.90	88.00
LINDNER	1.67	73.68	80.21	85.19	91.47
CHEN	1.17	86.67	92.67	95.54	98.53
ZHONG	1.12	86.91	91.82	94.08	97.90
Ours	1.07	88.14	93.68	96.03	97.60

In the test dataset, Ibragimov et al. and Lindner et al. achieved outstanding results by combining random forests and statistical shape models. Compared to their algorithms, CenterNet-PSA shows

significant improvements on the test dataset: the success detection rate (SDR) within a 2mm range increased by 16.44% and 14.46%, respectively, while the mean radial error (MRE) decreased by 0.77mm and 0.60mm, respectively. Compared to the state-of-the-art method of Zhong et al., the proposed algorithm in this study achieves higher success detection rates in the 2.0mm, 2.5mm, 3mm, and 4mm detection ranges. Furthermore, Zhong et al. used a multi-stage U-Net network and patch-based methods for heatmap regression. While these methods can improve accuracy, they significantly increase computational time and cost. In contrast, our method maintains high accuracy while being more efficient. In terms of the average radial error (MRE), the performance of our method on the test dataset is 1.07mm, surpassing other methods. Experimental results show that the proposed CenterNet-PSA model has powerful feature extraction capabilities and achieves excellent performance in landmark detection tasks. To provide a more intuitive demonstration of the model's effectiveness, we visualized some of the final prediction results on the test set, as shown in Figure 6. The red dots represent the 19 points annotated by the doctors, and the blue points represent the model's predictions. From the figure, it is clear that the predicted locations of our method closely match the true positions, which meets clinical needs effectively.

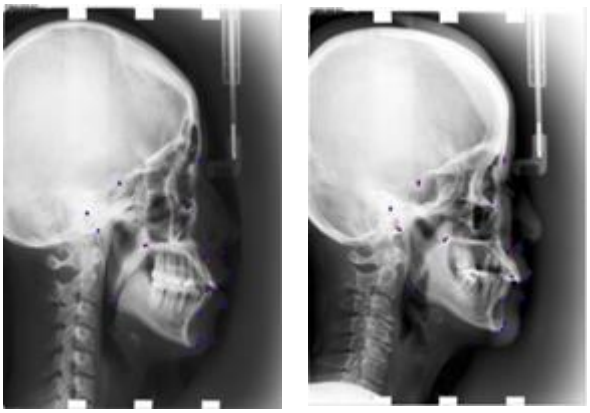

Figure 6. The visualize results

3.6. Ablation Study

This study validates the advantages of the proposed method through ablation experiments. Ablation experiments were conducted on the ISBI 2015 test dataset to evaluate the model's performance under different experimental settings. In the experiment, CenterNet was chosen as the base network, and the Polarized Self-Attention Mechanism was added on top of it. The experimental results are shown in Table 2. In the base network, the model's mean radial error (MRE) was 1.43mm, and the success detection rates within the localization error ranges of 2.0mm, 2.5mm, 3.0mm, and 4.0mm were 78.96%, 84.20%, 90.64%, and 94.25%, respectively. After adding the Polarized Self-Attention Mechanism to the base network, the model’s mean radial error decreased to 1.07mm, a reduction of 0.36mm compared to the base network. Additionally, within the same localization error ranges, the success detection rates increased to 88.14%, 93.68%, 96.03%, and 97.60%, respectively. These results demonstrate the clear advantages of the proposed method.

Table 2. Results of different models on Test

Method	MRE/mm	Success Detection Rate/%			
		Localization Error 2.0mm	Localization Error 2.5mm	Localization Error 3.0mm	Localization Error 4.0mm
CenterNet	1.43	78.96	84.20	90.64	94.25
CenterNet-PSA	1.07	88.14	93.68	96.03	97.60

4. CONCLUSION

In this study, we developed a novel network architecture called CenterNet-PSA, which leverages the Polarized Self-Attention (PSA) mechanism to accurately identify key landmarks in cephalometric measurements. The architecture integrates the PSA mechanism with the DLA-34 backbone, enhancing the model's feature learning ability, improving its overall perception of channel and spatial information, and reinforcing its ability to capture fine details. The model performs prediction tasks via heatmap regression. Experimental results demonstrate that our proposed algorithm excels in both MRE and SDR metrics, significantly improving accuracy while keeping errors within clinically acceptable limits, thus reducing clinicians' reliance on cephalometric analysis. This has high practical value in orthodontics. However, there are still areas for improvement. Currently, the research focuses primarily on detecting cephalometric landmarks in adults, whose cranial structure is clear, and dental arrangement is regular, overlooking the more challenging adolescent subjects. Adolescents may have mixed dentition and complex morphological changes, which can significantly displace cephalometric landmarks. Additionally, there is a lack of sufficient adolescent cephalometric X-ray data. In future research, we plan to focus on two areas:

Collecting adolescent cephalometric data to create a database;

Further improving the model to address cephalometric landmark detection across different age groups.

ACKNOWLEDGEMENTS

Shanxi Province Natural Science Foundation: Research on Core Scalable and Assurance Intelligent Service Model for Cross-Network Virtual Reality (201801D121147)

REFERENCES

[1] GRAU V, ALCANIZ M, JUAN M, et al. Automatic localization of cephalometric landmarks [J]. Journal of Biomedical Informatics, 2001, 34(3): 146-56.

[2] KEUSTERMANS J, MOLLEMANS W, VANDERMEULEN D, et al. Automated cephalometric landmark identification using shape and local appearance models; proceedings of the 2010 20th International Conference on Pattern Recognition, F, 2010 [C]. IEEE.

[3] IBRAGIMOV B, LIKAR B, PERNUS F, et al. Computerized cephalometry by game theory with shape-and appearance-based landmark refinement; proceedings of the Proceedings of International Symposium on Biomedical imaging (ISBI), F, 2015 [C].

[4] OKTAY O, BAI W, GUERRERO R, et al. Stratified decision forests for accurate anatomical landmark localization in cardiac images [J]. IEEE transactions on medical imaging, 2016, 36(1): 332-42.

[5] CRIMINISI A, SHOTTON J, BUCCIARELLI S. Decision forests with long-range spatial context for organ localization in CT volumes; proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), F, 2009 [C]. Citeseer.

[6] LINDNER C, COOTES T F. Fully automatic cephalometric evaluation using random forest regression-voting; proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI) 2015–Grand Challenges in Dental X-ray Image Analysis–Automated Detection and Analysis for Diagnosis in Cephalometric X-ray Image, F, 2015 [C].

[7] LEE H, PARK M, KIM J. Cephalometric landmark detection in dental x-ray images using convolutional neural networks; proceedings of the Medical imaging 2017: Computer-aided diagnosis, F, 2017 [C]. SPIE.

[8] ZHONG Z, LI J, ZHANG Z, et al. An attention-guided deep regression model for landmark detection in cephalograms; proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22, F, 2019 [C]. Springer.

[9] QIAN J, CHENG M, TAO Y, et al. CephaNet: An improved faster R-CNN for cephalometric landmark detection; proceedings of the 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), F, 2019 [C]. IEEE.

[10] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2014 [C].

[11] DAI X, ZHAO H, LIU T, et al. Locating anatomical landmarks on 2D lateral cephalograms through adversarial encoder-decoder networks [J]. IEEE Access, 2019, 7: 132738-47.

[12] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview [J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[13] LIU H, LIU F, FAN X, et al. Polarized self-attention: Towards high-quality pixel-wise regression [J]. arXiv preprint arXiv:210700782, 2021.

[14] ZHANG Q. A novel ResNet101 model based on dense dilated convolution for image classification [J]. SN Applied Sciences, 2022, 4: 1-13.

[15] SUSANTO Y, LIVINGSTONE A G, NG B C, et al. The hourglass model revisited [J]. IEEE Intelligent Systems, 2020, 35(5): 96-102.

[16] YU F, WANG D, SHELHAMER E, et al. Deep layer aggregation; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].