A Cephalometric Landmark Regression Method based on Dual-encoder for High-resolution X-ray Image

Chao Dai[1†], Yang Wang[2†(✉)], Chaolin Huang[3†], Jiakai Zhou[4], Qilin Xu[5], and Minpeng Xu[1]

[1] Tianjin University
[2] Anhui University of Technology
[3] Jiangxi University of Science and Technology
[4] Nanjing University of Aeronautics and Astronautics
[5] West Anhui University

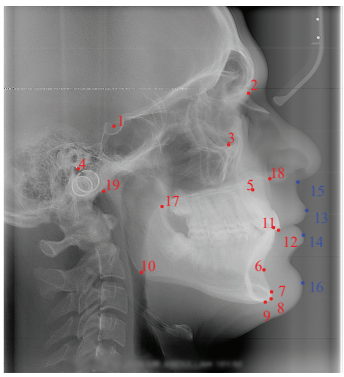
Abstract. Accurate detection of cephalometric landmarks is crucial for orthodontic diagnosis and treatment planning. Current methods rely on a cascading form of multiple models to achieve higher accuracy, which greatly complicates both training and deployment processes. In this paper, we introduce a novel regression paradigm capable of simultaneously detecting all cephalometric landmarks in high-resolution X-ray images. Our approach only utilizes the encoder module from the transformer to design a dual-encoder architecture, enabling precise detection of cephalometric landmark positions from coarse to fine. Specifically, the entire model architecture comprises three main components: a feature extractor module, a reference encoder module, and a finetune encoder module. These components are respectively responsible for feature extraction and fusion for X-ray images, coarse localization of cephalometric landmark, and fine-tuning of cephalometric landmark positioning. Notably, our framework is fully end-to-end differentiable and innately learns to exploit the interdependencies among cephalometric landmarks. Experiments demonstrate that our method significantly surpasses the current state-of-the-art methods in Mean Radical Error (MRE) and the 2mm Success Detection Rate (SDR) metrics, while also reducing computational resource consumption. The code is available at https://github.com/huang229/D-CeLR

Keywords: Cephalometric landmark · High-resolution · Dual-encoder · Reference encoder · Finetune encoder

1 Introduction

Cephalometric analysis represents a pivotal diagnostic tool extensively utilized in orthodontics and orthognathic surgery. This analysis involves the annotation of dental, skeletal, and soft tissue structures in lateral cephalometric radiographs.

† Equal contribution. (✉) Corresponding authors (youngnuaa@gmail.com).



19 Landmarks in Cephalometric			
1	Sella	11	Lower Incisor Tip
2	Nasion	12	Upper Incisor Tip
3	Orbitale	13	Labrale superius
4	Porion	14	Labrale inferius
5	Upper Incisor Apex	15	Subnasale
6	B-point	16	Soft Tissue Pogonion
7	Pogonion	17	Posterior Nasal Spine
8	Menton	18	Anterior Nasal Spine
9	Gnathion	19	Articulare
10	Gonion		

(a) Cephalometric landmark coordinate positions.

(b) Cephalometric landmark medical name

Fig. 1: Cephalometric landmark visualization. (a) Cephalometric landmark coordinate positions. Red indicates hard tissue points and blue indicates soft tissue points. (b) Cephalometric landmark medical name.

As illustrated in Figure 1, these cephalometric landmarks are core to the analysis, providing reference points for subsequent qualitative assessments of angles and distances. However, the manual annotation of these landmarks is a laborious, time-consuming, and highly subjective task, impacting the accuracy of the annotations. Consequently, a precise and robust automated method for annotating cephalometric landmarks holds significant importance for effective treatment planning [1, 6, 7, 14, 23].

Existing methods for cephalometric landmark detection can be broadly classified into two categories: heatmap-based and regression-based approaches. The heatmap-based approach involves predicting a heatmap that indicates the probability of each pixel in a region corresponding to various cephalometric landmarks. This modality has seen extensive applications in the detection of cephalometric landmarks. For example, Chen et al. [6] introduced a feature pyramid fusion-based heatmap method for simultaneous landmark detection, achieving impressive results. Qian J et al. [26] advanced the accuracy of cephalometric landmark detection by designing a multi-head attention module and a novel regional loss function. However, heatmap-based methods exhibit certain disadvantages. 1). The ground truth requires manual design and heuristic adjustments, with inevitable noise impacting the final outcomes [13, 29, 40]. 2). post-processing operations are necessary to locate single maximum values in heatmaps. These operations are typically heuristic and non-differentiable, undermining the model's capacity for end-to-end training. 3). models generally adopt a U-net structure [27, 28, 41], while processing high-resolution X-ray images, consumes more computational resources and is prone to missing cephalometric landmarks.

Regression-based methods directly map the input image to the coordinates of cephalometric landmarks, typically employing a feedforward network (FFN) for prediction. The regression-based methods is considerably more streamlined

compared to heatmap-based methods, as the prediction of cephalometric landmarks is inherently a process of determining a set of coordinate values. Numerous regression-based techniques exist for predicting cephalometric landmarks. For example, Song Y et al. [29] utilizes a base network for coarse localization of cephalometric landmarks, followed by region-specific cropping and refined positioning using a secondary model. Gilmour L et al. [11] constructs individual models for each landmark to predict their locations. Regression-based methods circumvent the necessity for non-maximum suppression, heatmap generation, and quantization error correction. However, to achieve higher precision on high-resolution X-ray images, current approaches predominantly rely on cascading multiple models, which compromises the inherent advantages of end-to-end training and prediction for regression-based methods.

To address these issues, we introduce a novel regression paradigm that exclusively utilizes the encoder module of transformer for the one-time detection of all cephalometric landmarks on high-resolution X-ray images. Specifically, we design a feature extraction module based on Convolutional Neural Networks (CNN) to accomplish feature extraction and fusion for X-ray images. Subsequently, the extracted features are fed into a reference encoder module for the coarse localization of cephalometric landmarks. Finally, the coarsely localized cephalometric landmarks, along with the fused features, are inputted into a finetune encoder module, which iteratively refines the positioning of the cephalometric landmarks from coarse to fine detail. Moreover, our method pioneers the complete end-to-end training and deployment for the detection of cephalometric landmarks on high-resolution X-ray images. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on popular benchmarks with a ResNet-34 backbone. Specifically, we achieve a Mean Radial Error (MRE) of 1.01mm, 1.27mm, and 0.9372mm on the ISBI2015 test1, ISBI2015 test2, and ISBI2023 test datasets, respectively. Furthermore, our method significantly reduces GFLOPs, by 132% compared to the previously best method [11].

The main contributions of this work are as follows:

- We propose an innovative regression paradigm for high-resolution X-ray images, which enables the prediction of all cephalometric landmarks through a single model. Moreover, our method facilitates end-to-end training and prediction, which not only improves efficiency but also enhances the feasibility of the model in practical applications.
- We have designed a dual-encoder structure, comprising a reference encoder module and a finetune encoder module. The reference encoder module accomplishes coarse localization of cephalometric landmarks, while the finetune encoder module refines this localization in a layer-by-layer updating manner.
- Our proposed regression approach significantly enhances the precision of cephalometric landmark detection. Compared to state-of-the-art methods, we achieve superior performance on both the ISBI2015 and ISBI2023 test datasets.

2 Related Work

With the seminal work of Lee et al. [19], which first introduced the use of deep learning for cephalometric landmark detection. Deep learning-based methods [2,17,34] have fully surpassed traditional pattern matching [4,10] and random forest regression-based methods [3, 22] in terms of accuracy for cephalometric landmark detection. This section primarily focuses on two deep learning-based approaches for cephalometric landmark detection and the transformer architectures for regression of keypoints.

2.1 Heatmap-Based Methods

Heatmap-based methods predict the likelihood of each pixel in the image corresponding to each cephalometric landmark. King C H et al. [17] utilized object detection techniques and designed a multitask loss without bounding box constraints to optimize landmark acquisition in the model. Chen R et al. [6] proposed a heatmap detection method based on feature pyramid fusion to complete all cephalometric landmark detection, surpassing other methods in effectiveness, but their multi-scale feature pyramid fusion is highly memory-intensive. Zhong Z et al. [40] adopted a two-stage landmark detection approach, which not only reduces memory consumption but also allows for fine-tuning of coarse landmark detection results on local image regions. Qian J et al. [26] enhanced the accuracy to new heights in the ISBI 2015 dataset by designing a multi-head attention module and a new regional loss function, while Ao Y et al. [2] developed a multiscale feature aggregation (MSFA) module and multi-head loss function. Although heatmap-based cephalometric landmark detection achieves high accuracy, its application to high-resolution X-ray images and the common use of U-net structures in models result in substantial memory resource consumption. Moreover, the post-processing required in heatmap-based methods disrupts the integrity of end-to-end training and deployment of the model.

2.2 Regression-Based Methods

Currently, the majority of regression methods for cephalometric landmark detection on high-resolution X-ray images utilize multi-stage or multi-model strategies. Song Yet et al. [29,30] proposed a method combining traditional regression algorithms with deep learning for coarse localization of landmarks, followed by cropping the region of interest in the original image to create a new image for refined localization using a secondary model. However, their accuracy is substantially lower than that achieved by heatmap-based methods [2, 26]. Zeng M et al. [36] introduced a three-tier cascading neural network for cephalometric landmark regression, akin to the concept used in the MTCNN model [38] for face detection. This approach significantly reduced memory resource consumption but did not achieve the desired level of accuracy. Gilmour L et al. trained 19 distinct models to predict each cephalometric landmark position, attaining accuracy on the ISBI 2015 cephalometric dataset comparable to heatmap-based

methods [26, 34]. This greatly encouraged the use of low-memory-consuming regression methods in landmark detection. However, the necessity of maintaining a separate model for each landmark adds complexity to training and deployment. While some regression methods have reached heatmap-based method accuracy levels, they typically involve designing multiple network models for predictions. Moreover, these methods have also not achieved end-to-end training and deployment.

2.3 Transformer-based architectures

The Transformer [31], proposed by Vaswani et al., originally designed for natural language processing tasks, employs an encoder-decoder architecture based on self-attention and feed-forward networks. Recently, Transformer-based models have demonstrated significant potential in computer vision tasks [5, 9], including various works applying the Transformer structure to keypoint estimation. Such as TransPose [33] and HRFormer [35] utilized the encoder-decoder structure of transformers for human keypoint regression. Poseur [24] and DTLD [20] have adopted the latest deformable transformer architecture for efficient regression of human keypoints and facial landmarks. Despite the high performance achieved by transformer-based methods in keypoint regression tasks, they present certain challenges: 1) They are primarily used for low-resolution images; 2) The deformable transformer architecture is more complex for deployment. In contrast, our method addresses these issues and achieves significantly higher performance.

3 Method

The overall architecture, as illustrated in Figure 2, presents our proposed dual-encoder model which progressively predicts cephalometric landmark coordinate from coarse to fine on high-resolution X-ray images. It comprises a feature extractor for image feature extraction, a reference encoder for coarse cephalometric landmark localization, and a finetune encoder for precise cephalometric landmark localization. For the input image, we initially obtain multi-scale features ($S2$, $S3$, $S4$, and $S5$) and a fused feature F_u through the feature extractor (Sec.3.1). The feature map $S5$ is flattened to produce the image context queries V_{FR}^{C}, and coarse landmark content queries V_{LR}^{C} are initialized randomly. The image context queries V_{FR}^{C} and coarse landmark context queries V_{LR}^{C} are fed into the reference encoder along with their position queries V_{R}^{P}, updating to corresponding context queries $V_{LR}^{C'}$ and $V_{FR}^{C'}$. Subsequently, the context queries $V_{LR}^{C'}$ are utilized to predict the coarse coordinate of cephalometric landmark $\mu_R \in R^{K\times 2}$ and coarse distribution $\sigma_R \in R^{K\times 1}$ via FFN (Sec.3.2). Next, the fused feature map Fu is also flattened to generate image context queries V_{FA}^{C}, and fine landmark content queries V_{LA}^{C} are initialized. Unlike the reference encoder module, which solely uses content and position queries as input, the coarse landmark coordinates μ_R and feature map F_u are also fed into the finetune encoder module to update the content queries $V_{LA}^{C'}$ and $V_{FA}^{C'}$. Finally, the content queries $V_{LA}^{C'}$ is operated

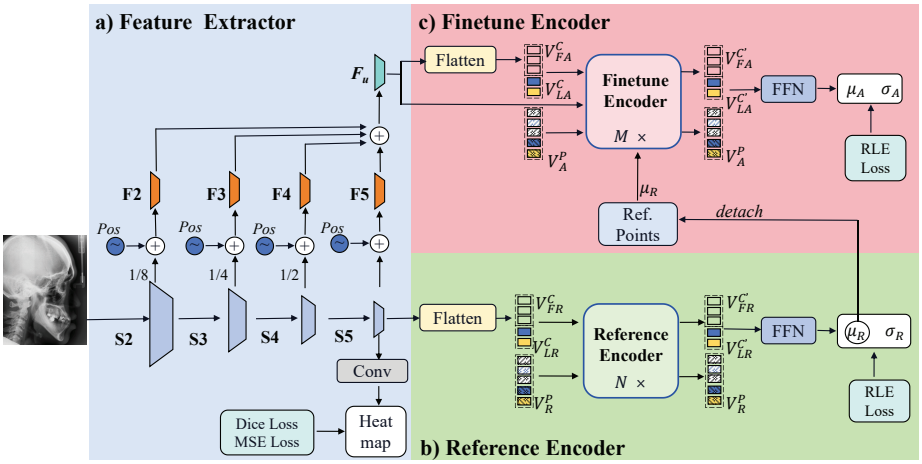


Fig. 2: The overview architecture of our method, which contains (a) feature extractor module, (b) reference encoder module and (c) finetune encoder module.

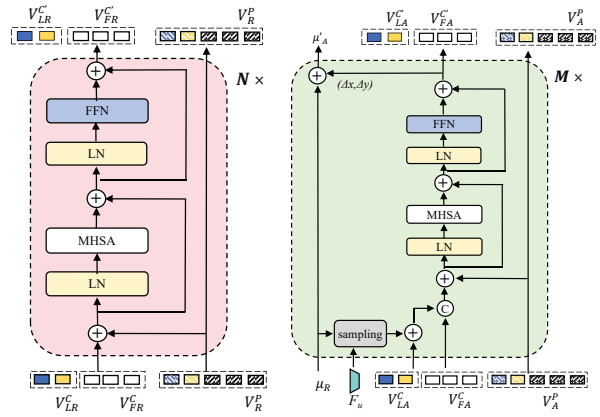
by the FFN to produce cephalometric landmark coordinate $\mu_A \in R^{K\times 2}$ and distribution $\sigma_A \in R^{K\times 1}$ (Sec.3.3). In addition, different loss functions are employed for supervising the training of various modules. For the feature extractor module, Dice loss and Mean Squared Error (MSE) loss are utilized to aid model optimization. For the reference encoder and finetune encoder modules, Residual Log-likelihood Estimation(RLE) loss is applied to optimize the model's output cephalometric landmark coordinates μ and distribution σ (Sec.3.4).

3.1 Feature Extractor

ResNet34 [16] is utilized as the backbone in our model, from which multi-level feature maps [39] are extracted, as illustrated in Figure 2. Initially, we apply downsampling operations to scale the feature maps $S2$, $S3$, and $S4$ to the same dimension and size as the feature map $S5$. Subsequently, the feature maps outputted by the backbone are summed with their respective positional maps (Pos) to yield new feature maps $F2$, $F3$, $F4$, and $F5$. These feature maps $F2$, $F3$, $F4$, and $F5$ are aggregated to generate the fused feature map F_u. The feature map $S5$ is directly fed into the reference encoder module to coarse locate cephalometric landmark, while the fused feature map F_u is fed into the finetune encoder module to precise locate cephalometric landmark. Moreover, to enhance the model's performance, the feature map $S5$ is processed through convolution to generate a heatmap, which is optimized by Dice loss and MSE loss.

3.2 Reference Encoder

The reference encoder aims to establish the relationship between cephalometric landmark queries and feature maps, thereby facilitating the coarse prediction of



(a) reference encoder module **(b)** finetune encoder module

Fig. 3: The detailed illustration of (a) reference encoder module and (b) finetune encoder module.

cephalometric landmark. As illustrated in Figures 2b and Figure 3a, the reference encoder module follows the typical transformer encoder paradigm. It comprises N identical layers within the encoder, each layer consisting of Layer Normalization (LN), Multi-Head Self-Attention (MHSA), and Feed-Forward Networks (FFN). Specifically, we initialize K cephalometric landmark content queries V_{CL}^{R} and utilize the feature map $S5$ as the image content queries V_{CF}^{R}. Drawing inspiration from the positional encoding of the BERT [8], we generate the positional queries V_{P}^{R}. These content and positional queries are fed into the reference encoder. After N layers of iteration, the reference encoder outputs the updated cephalometric landmark content queries $V_{LR}^{C'}$. These content queries are calculated by FFN layer to predict the coarse cephalometric landmark coordinates μ_R and distribution σ_R.

3.3 Finetune Encoder

The finetune encoder employs a layer-to-layer update mechanism to achieve more precise cephalometric landmark detection. The structure of the finetune encoder, as shown in Figure 2c and Figure 3b, also adheres to the typical transformer encoder paradigm, consisting of M identical layers within the encoder. Unlike the reference encoder module, cephalometric landmark coordinate μ_R is continually updated in each layer of the finetune encoder module. Specifically, we initialize K cephalometric landmark content queries V_{LA}^{C} and flatten the fused feature map F_u to serve as the image content queries V_{FA}^{C}. Drawing inspiration from the positional encoding of the BERT, we generate position queries V_{A}^{P}. Five parameters are fed into the finetune encoder module, namely fine landmark context queries V_{LA}^{C}, image context queries V_{FA}^{C}, position queries V_{A}^{P}, the fused feature map F_u,

and coarse landmark coordinates μ_R. Within the finetune encoder module, we first sample feature vectors on the fused feature map F_u using coarse cephalometric landmark coordinates μ_R, then add it to the fine landmark queries V_{LA}^{C}. We combine content and position queries and feed them into the encoder to calculate the relationships among fine landmark and image context queries. Next, to adjust the landmark positions, we use the updated cephalometric landmark content queries $V_{LA}^{C'}$ to calculate the $(\Delta x, \Delta y)$ offsets by FFN layer and add them back to the previous cephalometric landmark coordinates μ_R. In this way, the finetune encoder module refines the content queries progressively by stacking multiple aforementioned layers, outputting $V_{LA}^{C'}$ and $V_{FA}^{C'}$. Finally, the cephalometric landmark content queries $V_{LA}^{C'}$, followed by FFN layer, predicts the fine cephalometric landmark coordinates μ_A and distribution σ_A.

3.4 Loss Function

As shown in Figure 2, the loss function of our method is composed of two key components: 1) The heatmap loss of the feature extraction module, 2) The cephalometric landmark regression loss for both the reference encoder and finetune encoder modules. The overall loss function of our method can be formulated as follows:

$$L = \lambda_{HM} L_{HM} + \lambda_{RE} L_{RE} + \lambda_{FE} L_{FE} \tag{1}$$

where L_{HM} , L_{RE} and L_{FE} represent feature extraction, reference encoder, and finetune encoder module loss functions respectively. λ_{HM}, λ_{RE}, and, λ_{FE} are the hyper-parameters used to balance the three losses, and they are set to 1.0, 1.0, and 1.0, respectively. L_{HM} consists of the Dice loss and the MSE loss. L_{HM} is defined as follows:

$$L_{HM} = Dice(\hat{P}_{hp}, P_{hp}) + Mse(\hat{P}_{hp}, P_{hp}) \tag{2}$$

where $\hat{P}_{hp}$ and P_{hp} are the prediction heatmap and ground truth heatmap respectively. For the cephalometric landmark regression loss of the reference encoder module, we adopt Residual Log-likelihood Estimation(RLE) loss. The loss is defined as follows:

$$L_{RE} = RLE(\mu_R, \sigma_R; \mu_g) \tag{3}$$

where μ_R and σ_R are coarse cephalometric landmark coordinate and distribution output by the reference encoder module. μ_g is cephalometric landmark ground truth coordinate. For the cephalometric landmark regression loss of the finetune encoder module, we also adopt RLE loss. The loss is defined as follows:

$$L_{FE} = \sum_{i=1}^{M} RLE(\mu_{A,i}, \sigma_{A,i}; \mu_g) \tag{4}$$

where M is number of finetune encoder layer. $\mu_{A,i}$ and $\sigma_{A,i}$ represent the cephalometric landmark coordinate and distribution output by the i-th layer reference encoder module.

4 Experiments

In this section, we assess our method on some benchmarks for cephalometric landmark detection task. We first perform several ablation studies to underline the advantage of our proposed methods and to establish the optimal setting for hyperparameters. Finally, we compare the performance of our model with state-of-the-art methods.

4.1 Implementation Details

Our model is built on the PyTorch framework. We use ResNet-34, pre-trained on ImageNet, as the backbone. Our architecture includes 4 layers for both the reference encoder and finetune encoder module. All additional layers that we introduce are initialized randomly. The model training and testing are performed on one NVIDIA 3060(12GB) GPU. For model optimization, we use Adam [18], with parameters $\beta 1 = 0.9$, $\beta 2 = 0.999$, and a weight decay of 10^{-4}. The batch size is set to 4. The model is trained for 1000 epoch. The initial learning rate is 2×10^{-4}, and dynamically updated the learning rate using the cosine strategy during the training process. Data augmentation techniques are employed, encompassing random cropping and random rotation. For the random cropping operation, all cephalometric landmarks are preserved during each cropping process. Regarding the random rotation operation, we select a rotation angle range of [-30, 30] degrees. Ultimately, the image is scaled to 1024×1024 for both training and inference of the model.

4.2 Dataset and Evaluation Metric

ISBI 2015 Challenge Dataset [37]. This is a widely utilized benchmark dataset in the field of cephalometric landmark detection. This dataset comprises 400 cephalometric images, of which 150 are designated for training, 150 for Test 1, and the remaining images for Test 2. Each image has been annotated with 19 landmarks by two experienced medical practitioners, and the average of these annotations is taken as the ground truth. This dataset provides a rich array of annotated data, enabling researchers to effectively train and evaluate their cephalometric landmark detection methods.

ISBI 2023 Challenge Dataset [25]. This is a recently introduced cephalometric landmark detection dataset, collected from seven distinct imaging devices. Following the training strategy in reference [15], we randomly selected 500 images as training data, with the remaining 200 images utilized for evaluating model performance. Experiments were conducted with k-fold(k=10) method cross-validation, and the average results were considered as the final outcome. This dataset provides 29 landmarks, but only the same 19 landmarks as in the ISBI 2015 dataset are used, ensuring a fair comparison with other methods. This new dataset offers researchers a more challenging scenario to test the generalization capabilities of their methods across various imaging devices.

Evaluation metric. The evaluation of cephalometric landmark detection models typically employs the Mean Radial Error (MRE) and the Successful Detection Rate (SDR) [7]. MRE is used to calculate the distance error between the predicted cephalometric landmarks and the ground truth, commonly serving as a measure of detection accuracy. The calculation method for MRE is defined as follows:

$$R_i^j = \| \mu_A^j(x_i, y_i) - \mu_g^j(x_i, y_i) \|_2 \tag{5}$$

$$MRE = \frac{1}{TK} \sum_{i=1}^{T} \sum_{j=1}^{K} R_i^j \tag{6}$$

where R_i^j denotes the radial error of the $i-th$ landmark in the $j-th$ image. $\mu_A^j(x_i, y_i)$ represents the coordinates of the $i-th$ cephalometric landmark predicted for the j-th image. $\mu_g^j(x_i, y_i)$ denotes the ground truth coordinates of the$i-th$ cephalometric landmark in the $j-th$ image. T represents the number of test images, and K denotes the number of cephalometric landmark in each image. SDR is employed to quantify the discrepancy between the predicted cephalometric landmark and the ground-truth. If the radial error R_i^j is no greater than z mm (where z = 2mm, 2.5mm, 3mm, 4mm), the detection is considered as a successful one (Usually, 2mm range is acceptable in medical analysis [32, 40]). The SDR is defined as follows:

$$SDR_i = \frac{1}{TK} \sum_{j=1}^{T} \sum_{j=1}^{K} \{R_i^j < z\} \tag{7}$$

4.3 Ablation Study

In this section, we perform several ablation studies on ISBI 2015 Challenge dataset to illustrate the effectiveness of the proposed component.

Table 1: Varying different model structures. "MF" denotes Multi-level Features. "HP" denotes Heatmap. "RE" denotes Reference Encoder. "RL" denotes RLE Loss. "FE" denotes Finetune Encoder.

ID	Baseline	Feature Extractor module		Reference Encoder module		Finetune Encoder module		MRE(mm)	2mm(SDR%)
		MF	HP	RE	RL	FE	RL		
1	✓							2.8974	54.75
2	✓	✓						2.2586	61.91
3	✓		✓					2.5698	57.07
4	✓	✓	✓					2.0125	65.01
5	✓			✓				1.6745	74.04
6	✓			✓	✓			1.2434	83.65
7	✓	✓	✓	✓	✓			1.1468	86.84
8	✓	✓	✓			✓		1.1514	86.31
9	✓	✓	✓			✓	✓	1.0230	88.12
10	✓	✓	✓	✓	✓	✓	✓	1.0088	89.51

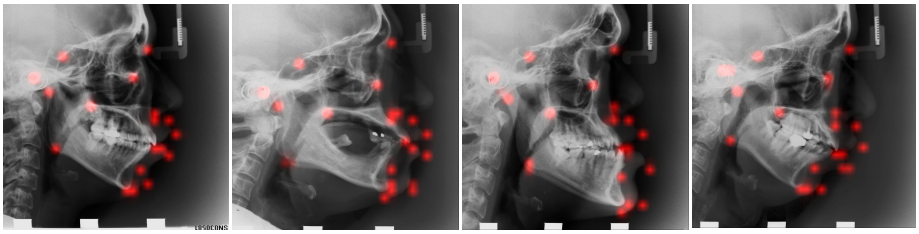


Fig. 4: Headmap visualization. The attention heatmap come from the feature extraction module.

Varying the model structures. We conduct experiments to verify the different model structures. All experimental results are presented in Table 1. Regarding the feature extractor module, the combination of the multi-level feature (MF) module improves the baseline in MRE and 2mm SDR indicators by 0.3276mm and 7.16% respectively, while the introduction of the heatmap (HP) module improves the baseline by 0.3276mm and 2.32%. When both MF and HP modules are integrated, there is 0.8849mm and 10.26% enhancement over the Baseline, underscoring the significant role of the feature extractor module in accuracy improvement. For the reference encoder module, the addition of reference encoder (RE) components and RLE Loss (RL) elements on the baseline foundation yielded 1.654mm and 28.9% accuracy improvement. When used in conjunction with the feature extractor module, the model's accuracy further increased by 0.0996mm and 3.19%. Regarding the finetune encoder module, its combined use with the feature extractor module led to a 1.8744mm and 33.37% improvement in model accuracy. The highest accuracy, reaching 1.0088mm and 89.51%, was achieved when the finetune encoder module was used in combination with both the reference encoder module and the feature extractor module. This underscores the significant impact of the three proposed modules on enhancing model accuracy. Finally, we visualize the attention heatmap in Figure 4. The heatmap is highly responsive at locations near the cephalometric landmarks.

Varying the levels of fuse feature map. We explore the impact of feeding different levels of fuse feature maps into the proposed finetune encoder. As shown in Table 2, the performance grows consistently with more levels of fuse feature maps, e.g., 89.20%, 89.33%, 89.42%, 89.51% for 2, 3, 4, 5 levels of feature maps on 2mm SDR, respectively.

Varying parameter of encoder module. We study the impact of encoder module on model performance from two aspects: the number of layers and feature dimensions. To simple the validation approach, the reference encoder and finetune encoder modules are set to the same number of layers and dimensions. First, we investigate the effects of altering the dimension of the encoder module. As illustrated in Table 3, there is a discernible enhancement in model efficacy concomitant with an increase in the dimensions of encoder layers. The peak performance of the model is attained when the dimension is augmented to 512. Furthermore, we conduct experiments by varying the number of encoder layers.

As shown in Table 4, the performance grows at the first four layers and saturates at the fifth decoder layer.

Table 2: Varying the scale levels of fuse feature map for feature extraction module.

F5	F4	F3	F2	MRE(mm)	SDR%			
					2mm	2.5mm	3mm	4mm
✓				1.0181	89.20	93.48	96.18	98.41
✓	✓			1.0156	89.33	93.49	96.32	98.41
✓	✓	✓		1.0113	89.42	93.50	96.38	98.54
✓	✓	✓	✓	1.0088	89.51	93.54	96.42	98.56

Table 3: Varying feature queries dimensions of encoder module.

Dim	MRE(mm)	SDR%			
		2mm	2.5mm	3mm	4mm
128	1.0201	89.03	93.26	95.76	98.17
256	1.0194	89.32	93.35	96.02	98.32
512	1.0088	89.51	93.58	96.42	98.56
768	1.0091	89.47	93.61	96.39	98.53

Varying the input image resolutions. We undertake experimental investigations to ascertain the robustness of our method across varying input resolutions. As depicted in Table 5, there is a significant enhancement in the performance of the model concomitant with an increase in the resolution of input images. When the input image resolution is 1024×1024, the model reaches 1.0088mm and 89.51% in MRE and 2mm SDR metrics respectively. A further escalation in input image resolution results in a decline for model performance.

Table 4: Varying the numbers of encoder layers.

Num	MRE(mm)	SDR%			
		2mm	2.5mm	3mm	4mm
1	1.0835	87.23	93.05	95.51	97.96
2	1.0247	88.98	93.31	95.92	98.32
3	1.0137	89.46	93.47	96.28	98.47
4	1.0088	89.51	93.54	96.42	98.56
5	1.0091	89.48	93.54	96.45	98.59

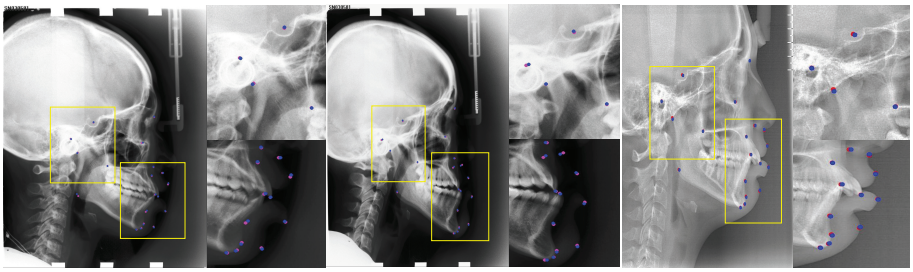
Table 5: Varying the input image resolutions.

Resolution	MRE(mm)	SDR%			
		2mm	2.5mm	3mm	4mm
256×256	1.2012	84.56	91.79	95.44	98.49
512×512	1.0674	88.07	93.30	96.25	98.57
768×768	1.0129	89.40	93.33	96.07	98.60
1024×1024	1.0088	89.51	93.54	96.42	98.56
1280×1280	1.0153	89.31	93.51	95.44	98.32

4.4 Main Result

We evaluated our method on two cephalometric landmark datasets: ISBI 2015 Challenge [37] and ISBI 2023 Challenge datasets [25]. The final results are presented in Tables 6,7,8. The proposed approach achieved the least Mean Radical Error (MRE) and the highest 2mm Success Detection Rate (SDR), which is considered as the clinically accepted. Moreover, our method achieves end-to-end training and prediction for cephalometric landmarks.

ISBI 2015 Challenge test1. Table 6 presents the evaluation results for the ISBI 2015 Challenge test1 dataset. These state-of-the-art methods can be categorized into heatmap-based and regression-based methods. Our method demonstrates clear superiority over heatmap-based methods. Compared to the best heatmap-based method [2], our method achieves improvements of 0.11mm and 1.48% respectively in MRE and the 2mm SDR metrics. Additionally, compared to the best regression-based method, our method achieves improvements of 1.19% on the 2mm SDR metrics. Moreover, compared to the best approach, our method exhibits a significant advantage in terms of GFLOPs. In addition, compared to other low-resolution methods, our method has the lowest GFLOPs of only 23.0, while the 2mm SDR reaches 88.07%, which is superior to the other



(a) ISBI 2015 Challenge test1 **(b)** ISBI 2015 Challenge test2 **(c)** ISBI 2023 Challenge

Fig. 5: Qualitative detection results on ISBI 2015 and 2023 Challenge datasets. (a) and (b) correspond the detection results for the ISBI 2015 Challenge test1 and test2. (c) depicts the detection outcomes for the ISBI 2023 Challenge. The blue landmarks represent results annotated by medical professionals, while the red landmarks indicate the outcomes predicted by the model.

Table 6: Quantitative results on the ISBI 2015 Challenge test1 dataset . * denotes other methods we implemented. Bold represents the best result.

Method	Backbone	Resolution	GFLOPs	MRE(mm)	SDR%			
					2mm	2.5mm	3mm	4mm
Heatmap-based Methods								
Chen R et al. [6]	ResNet50	800×640	215.7	1.17	86.67	92.67	95.54	98.53
Zhong Z et al. [40]	U-Net	290×290+19×100×100	92.2	1.12	86.91	91.82	94.88	97.90
CephaNN [26]	ResNeXt50	800×640	982.8	1.15	87.61	93.16	96.35	98.74
Yao J et al. [34]	ResNet18	576×512+19×96×96	40.1	1.14	86.84	93.02	95.43	**98.95**
Ao Y et al. [2]	Densenet121	800×640	157.2	1.12	88.03	92.73	95.96	98.48
Huang K et al. [13]	-	-	-	1.09	87.87	92.45	95.54	98.59
SimCC* [21]	HRNet48	800×640	164.9	1.12	87.16	91.96	95.37	98.18
Regression-based Methods								
Gilmour L et al. [11]	ResNet34	2432×1920	220.2	**1.01**	88.32	93.12	96.14	98.63
Song Y et al. [29]	ResNet50	256×256+19×256×256	102.5	1.08	86.40	91.70	94.80	97.80
Song Y et al. [30]	U-Net	480×387	286.8	1.19	85.20	91.20	94.40	97.20
Zeng M et al. [36]	-	-	-	1.34	81.37	89.09	93.79	97.86
King C H et al. [17]	-	-	-	1.17	86.14	91.72	94.91	97.96
Hong W et al. [12]	-	-	-	1.12	85.26	90.67	93.54	97.19
Poseur* [24]	ResNet50	800×640	46.1	1.14	86.56	91.09	94.00	97.23
Ours	ResNet34	512×512	23.0	1.07	88.07	93.30	96.25	98.57
Ours	ResNet34	1024×1024	95.0	**1.01**	**89.51**	**93.54**	**96.42**	98.56

methods. The qualitative detection results of the ISBI 2015 Challenge test1 dataset are displayed in Figure 4.

ISBI 2015 Challenge test2. The evaluation results for the ISBI 2015 Challenge test2 dataset are presented in Table 7. Our method outperforms heatmap-based methods by significant margins. Compared to best method [13], our method achieves an increase of 0.07mm in MRE and 0.48% in 2mm SDR. In addition, We introduce an end-to-end human keypoint detection method into the cephalometric landmark detection task, which is implemented based on the deformable decoder architecture. Experiments show that our method is significantly better than the human keypoint method in accuracy. Moreover, our

method is more convenient to deploy. Finally, the performance of the released methods on ISBI 2015 Challenge Test1 dataset are all better than Test2. It seems that the data distribution of Test1 dataset is more consistent with Train dataset. Qualitative detection results of our method on the ISBI 2015 Challenge test2 dataset can be found in Figure 5b.

Table 7: Quantitative results on the ISBI 2015 Challenge test2 dataset.

Method	MRE(mm)	SDR%			
		2mm	2.5mm	3mm	4mm
Heatmap-based Methods					
Chen R et al. [6]	1.48	75.05	82.84	88.53	95.05
Zhong Z et al. [40]	1.42	76.00	82.90	88.74	94.32
CephaNN [26]	1.43	76.32	82.95	87.95	94.63
Yao J et al. [34]	1.48	75.44	82.03	86.65	95.12
Ao Y et al. [2]	1.42	77.00	84.42	89.47	95.21
Huang K et al. [13]	1.34	79.05	**87.95**	89.79	95.05
SimCC* [21]	1.54	74.16	80.68	86.32	94.05
Regression-based Methods					
Gilmour L et al. [11]	1.33	77.05	83.16	88.84	94.89
Song Y et al. [29]	1.54	74.00	81.30	87.50	94.30
Song Y et al. [30]	1.64	72.20	79.50	85.00	93.50
Zeng M et al. [36]	1.64	70.58	79.53	86.05	93.32
King C H et al. [17]	1.50	74.58	81.74	87.26	94.73
Hong W et al. [12]	1.28	79.24	85.32	90.47	96.32
Poseur* [24]	1.48	74.42	81.37	86.68	93.63
Ours	**1.27**	**79.53**	86.47	**91.11**	**96.32**

Table 8: Quantitative results on the ISBI 2023 Challenge.

method	MRE(mm)	SDR%			
		2mm	2.5mm	3mm	4mm
Jin H et al. [15]	1.2200	83.76	89.71	92.79	96.08
Poseur* [24]	0.9982	88.51	92.82	95.37	97.79
SimCC* [21]	1.0795	88.39	93.12	95.31	97.81
Huang K et al.* [13]	1.0747	87.87	92.52	94.87	97.42
Gilmour L et al.* [11]	0.9793	89.37	93.47	95.97	97.42
Ours	**0.9372**	**90.68**	**94.24**	**95.97**	**97.89**

ISBI 2023 Challenge test. Regarding the ISBI 2023 Challenge test dataset, as illustrated in Table 8, Our method achieves the best performance on all metrics. Compared to the best-performing method [11], our approach significantly reduces the Mean Relative Error (MRE) from 0.9793mm to 0.9372mm and enhances the 2mm Success Detection Rate (SDR) from 89.37% to 90.68%. Moreover, in comparison with transformer-based methods, our approach demonstrates a lead of 0.061mm in MRE and 2.17% in 2mm SDR, respectively. Lastly, the qualitative detection results of our method on the ISBI 2023 Challenge test dataset are depicted in Figure 5c.

5 Conclusion

In this paper, we propose a novel regression model for cephalometric landmark detection for high-resolution X-ray image. This model only employs the encoder module within the transformer framework to construct the relationship between landmark features and image features. It is capable of regressing cephalometric landmark coordinate from coarse to fine and completes end-to-end training. Moreover, our model, compared to heatmap-based method, boasts low memory consumption and robustness against missing landmark. It offers a more straightforward end-to-end design compared to current regression-based method, performing one-time landmark detection on high-resolution X-ray images. Extensive experiments on the ISBI2015 and ISBI2023 datasets demonstrate that our method can achieve state-of-the-art performance compare with regression-based and heatmap-based methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62122059, 82330064).

References

1. Albarakati, S., Kula, K., Ghoneima, A.: The reliability and reproducibility of cephalometric measurements: a comparison of conventional and digital methods. Dentomaxillofacial Radiology **41**(1), 11–17 (2012)
2. Ao Y, W.H.: Feature aggregation and refinement network for 2d anatomical landmark detection. Journal of Digital Imaging **36**(2), 547–561 (2023)
3. B. Ibragimov, B. Likar, F.P., Vrtovec, T.: Automatic cephalometric x-ray landmark detection by applying game theory and random forests. In Proc. ISBI Int. Symp. on Biomedical Imaging (2014)
4. Cardillo, J., Sid-Ahmed, M.A.: An image processing system for locating craniofacial landmarks. IEEE transactions on medical imaging **13**(2), 275–289 (1994)
5. Carion N, Massa F, S.G.e.a.: End-to-end object detection with transformers. European conference on computer vision. Cham: Springer International Publishing pp. 213–229 (2020)
6. Chen, R., Ma, Y., Chen, N., Lee, D., Wang, W.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 873–881. Springer (2019)
7. Devereux, L., Moles, D., Cunningham, S.J., McKnight, M.: How important are lateral cephalometric radiographs in orthodontic treatment planning? American Journal of Orthodontics and Dentofacial Orthopedics **139**(2), e175–e181 (2011)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dosovitskiy A, Beyer L, K.A.e.a.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv preprint arXiv:2010.11929 (2020)
10. El-Feghi, M.S.A., Ahmadi, M.: Automatic localization of craniofacial landmarks for assisted cephalometry. Pattern Recognition **37**(3), 609–621 (2004)
11. Gilmour L, R.N.: Locating cephalometric x-ray landmarks with foveated pyramid attention. Medical Imaging With Deep Learning. PMLR pp. 262–276 (2020)
12. Hong W, Kim S M, C.J.e.a.: Deep reinforcement learning using a multi-scale agent with a normalized reward strategy for automatic cephalometric landmark detection. 2023 4th International Conference on Big Data Analytics and Practices pp. 1–6 (2023)
13. Huang K, F.F.: An intelligent shooting reward learning network scheme for medical image landmark detection. Applied Sciences **12**(20), 10190 (2022)
14. Indermun S, Shaik S, N.C.J.K.M.R.: Human examination and artificial intelligence in cephalometric landmark detection—is ai ready to take over? Dentomaxillofac Radiol **10.1259/dmfr.20220362** (2023)
15. Jin H, Che H, C.H.: Unsupervised domain adaptation for anatomical landmark detection. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland pp. 695–705 (2023)

16. Kaiming He, Xiangyu Zhang, S.R.J.S.: Deep residual learning for image recognition. ArXiv preprint arXiv:1512.03385 (2015)
17. King C H, Wang Y L, L.W.Y.e.a.: Automatic cephalometric landmark detection on x-ray images using object detection. 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) pp. 1–4 (2022)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980 (2014)
19. Lee H, Park M, K.J.: Cephalometric landmark detection in dental x-ray images using convolutional neural networks. Medical imaging 2017: Computer-aided diagnosis **10134**, 494–499 (2017)
20. Li, H., Guo, Z., Rhee, S.M., Han, S., Han, J.J.: Towards accurate facial landmark detection via cascaded transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4176–4185 (2022)
21. Li Y, Yang S, L.P.e.a.: Simcc: A simple coordinate classification perspective for human pose estimation. European Conference on Computer Vision. Cham: Springer Nature Switzerland **89-106** (2022)
22. Lindner, C., Cootes, T.: Fully automatic cephalometric evaluation using random forest regression-voting. IEEE International Symposium on Biomedical Imaging (ISBI) (2015)
23. Mamta Juneja, Poojita Garg, R.K.e.a.: A review on cephalometric landmark detection techniques. Biomedical Signal Processing and Control **66**(102486) (2021)
24. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., Hengel, A.v.d.: Poseur: Direct human pose regression with transformers. Proceedings of the European Conference on Computer Vision (ECCV) (October 2022)
25. Muhammad Anwaar Khalid, K.Z.e.a.: Cepha29: Automatic cephalometric landmark detection challenge 2023. ArXiv preprint arXiv:2212.04808 (2022)
26. Qian J, Luo W, C.M.e.a.: Cephann: a multi-head attention network for cephalometric landmark detection. IEEE Access **8**, 112633–112641 (2020)
27. Ronneberger O, Fischer P, B.T.: U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany,October pp. 5–9
28. Shaker A, Maaz M, R.H.e.a.: Unetr++: delving into efficient and accurate 3d medical image segmentation. ArXiv preprint arXiv:2212.04497 (2022)
29. Song, Y., Qiao, X., Iwamoto, Y., Chen, Y.w.: Automatic cephalometric landmark detection on x-ray images using a deep-learning method. Applied Sciences **10**(7), 2547 (2020)
30. Song Y, Qiao X, I.Y.e.a.: An efficient deep learning based coarse-to-fine cephalometric landmark detection method. IEICE TRANSACTIONS on Information and Systems **104**(8), 1359–1366 (2021)
31. Vaswani A, Shazeer N, P.N.e.a.: Attention is all you need. Advances in neural information processing systems (2017)
32. Wang, C.W., Huang, C.T., Hsieh, M.C., Li, C.H., Chang, S.W., Li, W.C., Vandaele, R., Marée, R., Jodogne, S., Geurts, P., et al.: Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. IEEE transactions on medical imaging **34**(9), 1890–1900 (2015)
33. Yang, S., Quan, Z., Nie, M., Yang, W.: Transpose: Keypoint localization via transformer. IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
34. Yao J, Zeng W, H.T.e.a.: Automatic localization of cephalometric landmarks based on convolutional neural network. American journal of orthodontics and dentofacial orthopedics **161**(3), e250–e259 (2022)

35. Yuhui Yuan, Rao Fu, L.H.W.L.C.Z.X.C.J.W.: Hrformer: High-resolution transformer for dense prediction. ArXiv preprint arXiv:2110.09408 (2021)
36. Zeng M, Yan Z, L.S.e.a.: Cascaded convolutional networks for automatic cephalometric landmark detection. Medical Image Analysis **68**, 101904 (2021)
37. Zhang H, Zhang J, L.C.S.E.S.P.N.T.G.S.W.Y.M.M.: All-net: Anatomical information lesion-wise loss function integrated into neural network for multiple sclerosis lesion segmentation. Neuroimage Clin **32**(102854) (2021)
38. Zhang K, Zhang Z, L.Z.e.a.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)
39. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. CVPR (2019)
40. Zhong Z, Li J, Z.Z.e.a.: An attention-guided deep regression model for landmark detection in cephalograms. Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China p. 13–17 (October 2019)
41. Ziyang Ye, H.Y., Li, B.: Uncertainty-aware u-net for medical landmark detection. Arxiv preprint arXiv:2303.10349v1 (2023)