GeoRefineNet: A Multistage Framework for Enhanced Cephalometric Landmark Detection in CBCT Images Using 3D Geometric Information

Thanaporn Viriyasaranon[1,3], Serie Ma[2], and Jang-Hwan Choi[1,3]

[1] Department of Computational Medicine, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, South Korea
[2] Division of Mechanical and Biomedical Engineering, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, South Korea
[3] Department of Artificial Intelligence, Ewha Womans University, Seoul, South Korea
thanaporn.v@ewhain.net, serie@ewhain.net, choij@ewha.ac.kr

Abstract. The precise detection of cephalometric landmarks on two-dimensional (2D) radiographs or three-dimensional (3D) computed tomography (CT) images is a fundamental step in various medical fields, especially in research on orthodontics and maxillofacial surgery. Deep learning-based detectors have demonstrated remarkable accuracy in 2D cephalometric analysis, whereas conventional single-view approaches are limited by their reliance on information from a single perspective. This study proposes GeoRefineNet, a novel multistage framework that leverages information from multiple CT scans acquired at various angles. By incorporating geometric knowledge through a 3D heatmap reconstruction process, GeoRefineNet improves robustness, accuracy, and adaptability to various cephalometric configurations. The proposed framework predicts 3D landmark positions on CT images, effectively addressing challenges associated with high-dimensional input data and limited training examples. GeoRefineNet surpasses the existing state-of-the-art models in the 2D and 3D domains, as demonstrated by its superior performance on numerical and clinical datasets. These findings indicate that GeoRefineNet offers a promising avenue for improving the accuracy and reliability of cephalometric landmark detection fostering further advances in clinical diagnosis and treatment planning. Our code is available at https://github.com/Thanaporn09/GeoRefineNet.git.

Keywords: Cephalometric landmark detection · Cone-Beam CT · Heatmap reconstruction · Multistage deep learning framework

1 Introduction

The precise identification of cephalometric landmarks on two-dimensional (2D) radiographs or three-dimensional (3D) computed tomography (CT) images is pivotal for clinical practice, such as maxillofacial surgery and orthodontics. Further, CT imaging provides a more detailed visualization of the craniofacial

area than traditional cephalometric radiographs. It also enhances the accuracy of landmark identification crucial for diagnosis, planning treatments, and evaluating patient outcomes [8,13].

Deep learning technologies have demonstrated impressive success in cephalometric analysis, particularly in detecting 2D cephalometric landmarks [12,16,17,19,28], especially within the domain of X-ray imaging [1], which contains only a single view of each patient. Consequently, existing methods have been designed to achieve high accuracy on single-view images, limiting their effectiveness in multiview landmark detection applications. Single-view approaches often fail to capture complex 3D anatomical relationships and the contextual information regarding surrounding structures. Recognizing these limitations, there is a growing interest in leveraging multiview information to improve detection accuracy.

In CT landmark detection, multiview information is particularly advantageous, providing additional data and increasing the robustness of detection methods against variations in anatomy, patient positioning, and imaging conditions. In some 2D projection views, landmarks might be obscured by the surrounding anatomy, but they can be identified in other views. Using clearer views to inform detection tasks in more challenging views can enhance the overall accuracy and reduce false positives and negatives. Moreover, a multiview analysis offers better handling of occlusions, making landmark detection more reliable. Although multi-view consistency has been shown to improve performance in spinal X-ray landmark detection [29], this approach increases computational costs and limits scalability due to its reliance on fusing feature representations from each view. To address these limitations, we propose GeoRefineNet, a novel framework for cephalometric landmark detection. GeoRefineNet leverages geometric information from multiple CT scans acquired from diverse angles, via a 3D heatmap reconstruction process, to enhance the robustness, accuracy, and adaptability of deep learning detectors without incurring additional model complexity or computational cost.

Accurate 3D cephalometric landmark detection remains a challenge due to the scarcity of robust methodologies and is hampered by high-dimensional input data and limited training examples. Although the existing methods [4,10,12,15,27] leverage the entire 3D CT volume, this approach inherently restricts the ability to overcome these limitations. We propose a novel framework that predicts 3D landmark positions directly on CT images by back-projecting the 2D output's landmark position from multiple angles acquired from GeoRefineNet. This approach offers two critical advantages: 1) accurate landmark detection without incurring the computational cost of high-dimensional input and 2) reduced overfitting risk due to the limited size of the training dataset.

The primary contributions of this paper are as follows:

– GeoRefineNet, an innovative cephalometric landmark detection framework that leverages geometric information across multiple views, has been introduced. This method is achieved by projecting 2D position onto a 3D heatmap in a reconstruction process, bridging the gap between 2D projections and 3D spatial detection.

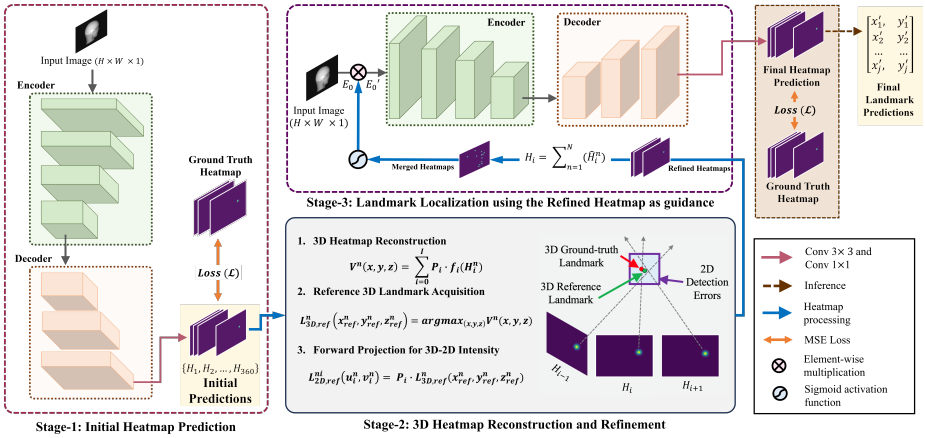

Fig. 1: Overall architecture of the proposed framework, GeoRefineNet.

- The integration of GeoRefineNet with 3D reconstruction processes facilitates the precise detection of landmarks in 3D space.
- GeoRefineNet surpasses existing models in accuracy across 2D projection and 3D imaging domains. This performance is validated using two CT image datasets: the 4D extended cardiac-torso (XCAT) head phantom dataset and patient CT scans from the CQ500 dataset [5].

2 Methods

This section introduces GeoRefineNet, a multistage deep learning framework for cone-beam CT (CBCT) cephalometric landmark detection that utilizes geometric information to refine landmark positions. GeoRefineNet consists of three primary stages: initial heatmap prediction, 3D heatmap reconstruction and refinement, and landmark localization using the refined heatmap as guidance. Figure 1 presents the overall architecture of GeoRefineNet.

2.1 Stage 1: Initial heatmap prediction

In the initial stage, we employed a deep learning anatomical landmark detector based on heatmap regression approaches to generate the initial heatmap, representing the predicted positions of landmarks on the 2D projection images of the CBCT scans. Typically, a heatmap regression-based detector comprises two main components: the encoder and the decoder. The encoder extracts the feature representation from the input image and generally serves as the backbone of the classification model, without the fully connected layer, in the natural image domain landmark detector. The decoder utilizes the feature representation from the encoder to generate the heatmap that corresponds to the landmark

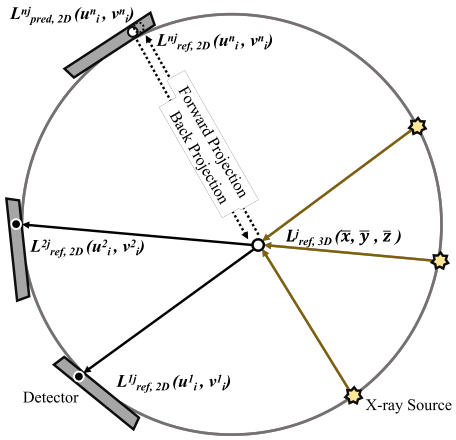

Fig. 2: Two- and three-dimensional transformation process..

positions. In medical imaging, heatmap regression-based landmark detectors are typically designed based on the UNet architecture [20].

The standard label representation for the heatmap regression-based method is a coordinate heatmap, generated as a 2-dimensional Gaussian distribution/kernel centered at the labeled coordinate of each landmark. In this study, the standard deviation of the Gaussian distribution is defined as two. The loss function employed is the mean square error (MSE) or L2 loss, calculated using the coordinate heatmap ground truth and the predicted heatmap. The predicted heatmap from the landmark detector serves as the initial heatmap for the subsequent stage.

2.2 Stage 2: 3D heatmap reconstruction and refinement

To mitigate the predicted errors in the first stage, we first reconstructed the initial heatmaps from the previous stage in a 3D space, using CBCT geometric information. Subsequently, we determined the positions of the 3D reference landmarks using 3D heatmap information. Finally, we forward-projected these positions onto a 2D space and encoded them as 2D Gaussian kernels, generating refined 2D heatmaps as guidance for the final landmark localization.

The projection matrix $P_i \in \mathbb{R}^{3\times4}$ plays a fundamental role in both the forward projection (3D to 2D) and the reconstruction (2D to 3D) processes. The projection matrix can be decomposed into extrinsic and intrinsic parts. The extrinsic part includes the translation $t_i \in \mathbb{R}^{3\times1}$ and the rotation $R_i \in \mathbb{R}^{3\times3}$ of the camera center in the world coordinate system. The intrinsic matrix $K_i \in \mathbb{R}^{3\times3}$ describes the mapping from the 3D camera coordinates to the 2D pixel coordinates. The decomposition of P_i is expressed as follows:

$$P_i = K_i \cdot [R_i|t_i] \tag{1}$$

To reconstruct the 3D heatmap from the initial 2D heatmaps obtained at each projection angle, we employed a 2D filtered backprojection algorithm [9] that leverages the known cone-beam computed tomography (CBCT) geometry. The filtered backprojection algorithm consists of three main steps: (1) pre-scaling the projections using cosine weighting to account for the varying distances between the X-ray source and detector pixels; (2) applying row-wise ramp filtering to the pre-scaled data using the Fourier transform of the Ram-Lak filter to enhance high-frequency components; and (3) backprojecting the ramp-filtered projections into the 3D volume using the known CBCT geometry.

The reconstructed 3D heatmap for the $n{-}th$ landmark, denoted as $V^n(x, y, z)$, is obtained by aggregating the filtered and backprojected initial heatmaps H_i^n from all projection angles $i \in [1, ..., I]$, where I is set to 360 in our implementation:

$$V^n(x, y, z) = \sum_{i=0}^{I} B_i \cdot f_i(H_i^n) \tag{2}$$

where f_i represents the filtering function applied during the ramp filtering step, and B_i is the backprojection operator, which reverse the projection process by distributing the filtered 2D projection data back into the 3D volume along the paths defined by P_i

Typically, the predicted landmark positions in each 2D projection image include detection errors stemming from ambiguity in landmark features or limitations in model capacity. In this study, we mitigated individual image detection errors by leveraging the landmark positions from multiple 2D projection images of the same patients, estimating the 3D reference landmarks via the 3D heatmap volume, as demonstrated in Figure 3. The coordinates of the highest intensity position in the 3D heatmap volume of each landmark, denoted as $L^n_{3D,ref}(x^n_{ref}, y^n_{ref}, z^n_{ref})$, are defined as the 3D reference position for each landmark as follows:

$$L^n_{3D,ref}(x^n_{ref}, y^n_{ref}, z^n_{ref}) = argmax_{(x,y,z)} V^n(x, y, z) \tag{3}$$

The 3D reference positions of various landmarks were individually estimated. Subsequently, each of these reference landmarks n underwent forward projection onto the projections, resulting in 2D reference positions denoted as $L^{ni}_{2D,ref}(u^n_i, v^n_i)$. Mapping the 3D reference landmarks onto the i-th projection image in 2D is expressed in Equation 4 using homogeneous coordinates:

$$L^{ni}_{2D,ref}(u^n_i, v^n_i) = P_i \cdot L^n_{3D,ref}(x^n_{ref}, y^n_{ref}, z^n_{ref}) \tag{4}$$

Finally, we generated the refined 2D heatmap, denoted as $\hat{H}^n_i$, by encoding 2D reference landmark coordinates as 2D Gaussian kernels with a standard deviation (sigma) of two for all landmarks.

2.3 Stage 3: Landmark localization using the refined heatmap as guidance

To enhance the detectability of deep learning-based landmark detectors, we utilized refined heatmaps as attention maps, incorporating prior landmark

position-related information to guide the detectors. These attention maps provide focused regional or spatial information related to landmark positions, thereby improving the feature extraction and localization capabilities of the detectors.

The refined heatmap $\hat{H}_i^n$ has dimensions of $h \times w \times N$, where each channel represents the heatmap for a specific landmark n. To generate the attention map A_i for each 2D projection image, we aggregated the heatmaps of each landmark into a single comprehensive heatmap, defined as the merged heatmap. Subsequently, the merged heatmap undergoes the sigmoid activation function, denoted by σ, to compute the attention scores, transforming the significance of features into a probabilistic distribution [7]. Furthermore, applying the sigmoid activation function to normalize the merged heatmap before multiplying it with the input image aids in managing gradient flow. The attention map generation process is presented in Equation 5.

$$A_i = \sigma(\sum_{n=1}^{N} \hat{H}_i^n) \tag{5}$$

Afterward, the attention map is applied to the 2D projection images through elementwise multiplication, serving as the input images for landmark localization, as depicted in Figure 1. In this stage, the attention map scales the pixel values of the input images instead of zeroing them out, ensuring that gradients can flow back through the network to prevent gradient vanishing. This method provides spatial information related to landmark positions without increasing the model's complexity or computational requirements.

2.4 3D detection using 3D heatmap reconstruction

To determine the landmark position in 3D space, the final 2D coordinates from the last stage of GeoRefineNet are back-projected into 3D space, estimating the 3D landmarks' positions. This process mirrors the procedure used in stage 2 of GeoRefineNet.

3 Experiments

3.1 Dataset

To evaluate the performance of GeoRefineNet, we conducted experiments on two CBCT datasets, including the XCAT phantom CT dataset and publicly accessible actual patient CT volumes from the CQ500 dataset [5]. For the XCAT CT dataset, head models were generated from the 4D XCAT phantom CT dataset [21] for 27 patients with varying anatomical sizes and genders. We manually labeled 11 cephalometric landmarks on the CT phantom volumes. The average resolution of the CT volume is $800 \times 800 \times 248$. Moreover, the isotropic voxel spacing is $0.5 \times 0.5 \times 1$ mm.

The publicly available CQ500 dataset consists of CBCT scans depicting patients with diverse medical conditions. This dataset includes scans that highlight

deformed shapes of patients' heads, presenting a considerable challenge for labeling landmark positions. Therefore, we specifically chose CBCT scans from a subgroup of 22 patients identified as normal, ensuring that the scans cover areas from the crown of the head to the upper teeth. Each CT volume was manually annotated with 10 landmarks. The average resolution of the CT volume is 512 $\times$ 512 $\times$ 293, with varying isotropic voxel spacing across patients.

To ensure the reliability of the quality of manual landmark labeling by following a standardized procedure, as outlined in a previous study [11]. Specifically, two annotators independently labeled the landmarks using the same set of detailed instructions. To assess consistency between the annotators, we calculated the inter-annotator agreement using Cohen's kappa, which resulted in a score of 0.83, indicating strong agreement.

We conducted forward projection on the 3D CT volumes and landmarks at 360 angles per patient to acquire 2D images and landmark labels for the 2D cephalometric landmark detection task on the XCAT CT and CQ500 datasets. The total number of images for 2D landmark detection is 9,720 and 7,920 for the XCAT and CQ500 datasets, respectively. Each 2D projection in XCAT has dimensions of 620 $\times$ 480 pixels with 0.616 mm pixel spacing, while in CQ500, it has dimensions of 800 $\times$ 600 pixels with 0.638 mm pixel spacing.

3.2 Implementation details

In the experiments, we implemented the proposed framework using MMPose [6], an open-source toolbox for pose estimation based on PyTorch, for the 2D cephalometric landmark detection task. Additionally, we performed three-fold cross-validation for all experiments in this study. For 2D landmark detection on the XCAT CT dataset, the framework was trained with an initial learning rate set to 0.0004, and the input images were resized to 1216 $\times$ 960 pixels. For 2D landmark detection on the CQ500 dataset, the initial learning rate was set to 0.0005, with an input image size of 800 $\times$ 600 pixels. The framework was trained for 100 epochs on both datasets using the AdamW optimizer.

Furthermore, we implemented the reconstruction projector using the PYRO-NN library [23], an open-source library for reconstruction operations, to generate the 3D heatmap volume in stage 2. In the reconstruction process, we adopted the Ram-Lak filter in the Fourier domain. For all experiments, the evaluation metrics included the mean radial error (MRE, mm) and the successful detection rate (SDR,%) under 2, 2.5, 3, and 4 mm conditions. MRE is the average of absolute euclidean error distance between the ground truth and predicted landmarks, and can be described as

$$R = \sqrt{\Delta x^2 + \Delta y^2}, \quad MRE = \frac{\sum_{i=1}^{N} R_i}{N}, \tag{6}$$

where N indicates the number of landmarks. SDR is calculated as the percentage of successfully predicted landmark coordinates within ranges of 2 mm, 2.5 mm,

3 mm, and 4 mm, and is formulated as

$$SDR = \frac{\#\ accurate\ detections}{\#\ total\ detections} \times 100\% \tag{7}$$

3.3 Performance evaluation

This section quantitatively compares GeoRefineNet with the state-of-the-art methods for 2D and 3D cephalometric landmark detection tasks. Additionally, we conducted ablation studies on the proposed framework to evaluate the contributions of its components. The metrics indicating the best performance in each table below are in bold font.

Comparisons with state-of-the-art methods: We evaluated the performance of GeoRefineNet in 2D cephalometric landmark detection using 2D projection images from the XCAT CT and CQ500 datasets. Table 1 presents the performance comparison of the proposed method with previous state-of-the-art methods on the XCAT CT and CQ500 datasets for the 2D cephalometric landmark detection task.

For the XCAT CT dataset, the GeoRefineNet combined with HTC and Multiresolution learning demonstrates significant improvements over other methods. The MRE for GeoRefineNet with HTC and Multiresolution is 1.61 mm with a standard deviation of 1.47 mm, which is the lowest among all compared methods. The SDRs at 2 mm, 2.5 mm, 3 mm, and 4 mm are 75.76%, 84.6%, 89.72%, and 95.06%, respectively. These values indicate a substantial enhancement in detection accuracy, especially when compared to standalone models such as Hourglass, HRNet-W48, and UNet, which exhibit lower SDRs and higher MREs.

In the CQ500 dataset, the GeoRefineNet combined with HTC and Multiresolution learning also outperforms other methods. The MRE achieved is 4.58 mm with a standard deviation of 3.56 mm. The SDRs for 2 mm, 2.5 mm, 3 mm, and 4 mm are 25.61%, 35.84%, 45.38%, and 60.39%, respectively. This performance

Table 1: Performance comparison of the proposed method with previous state-of-the-art methods on the XCAT CT and CQ500 datasets on a 2D task.

Model	#Param(M)	XCAT CT dataset					CQ500 dataset				
		MRE(SD)↓	SDR(%)↑				MRE(SD)↓	SDR(%)↑			
			2mm	2.5mm	3mm	4mm		2mm	2.5mm	3mm	4mm
Natural-domain landmark detectors											
Hourglass [14]	94.85	2.63 (2.66)	48.01	61.48	71.94	84.69	7.88 (10.92)	11.62	16.26	22.20	34.23
HRNet-W48 [22]	65.33	2.96 (3.82)	43.27	56.14	66.74	81.04	7.15 (9.03)	12.12	18.05	24.53	38.16
HRFormer-S [26]	44.04	2.58 (2.64)	48.32	62.12	72.60	85.60	6.39 (6.63)	13.34	19.75	26.40	40.61
UNet [20]	35.35	2.80 (4.88)	49.35	62.50	72.29	84.81	9.62 (18.01)	14.18	20.75	27.52	40.50
PVT-Tiny [25]	16.91	3.33 (4.07)	37.16	49.54	60.48	76.48	7.75 (8.00)	11.35	16.92	22.94	34.37
Conformer-Ti [18]	22.32	3.05 (3.51)	41.07	53.79	64.52	79.43	8.90 (10.13)	9.81	14.53	26.18	30.58
Medical-domain landmark detectors											
GU2Net [31]	2.74	3.64 (5.02)	37.09	48.86	58.42	73.20	11.19 (17.60)	10.19	18.20	19.90	30.25
AFPF [3]	78.97	3.14 (8.99)	56.83	68.85	77.47	87.54	14.65 (30.524)	12.90	18.73	24.50	35.41
FARNet [2]	20.68	2.91 (3.44)	42.97	55.50	66.55	80.90	8.03 (13.97)	9.25	14.29	20.06	32.60
GeoRefineNet - FARNet	20.68	2.27 (1.63)	53.47	67.57	77.48	88.93	6.23 (6.21)	13.78	20.72	26.83	40.97
HTC+Multiresolution learning [24]	16.20	1.93 (2.27)	66.74	77.30	84.19	91.82	5.24 (5.39)	20.05	28.41	36.76	51.97
GeoRefineNet - HTC+Multiresolution	16.20	**1.61 (1.47)**	**75.76**	**84.62**	**89.72**	**95.06**	**4.58 (3.56)**	**25.61**	**35.84**	**45.38**	**60.39**

Table 2: Performance comparison of the proposed method with previous state-of-the-art methods on the XCAT CT and CQ500 datasets on a 3D task.

Model	#Param(M)	XCAT CT dataset					CQ500 dataset				
		MRE(SD)↓	SDR(%)↑				MRE(SD)↓	SDR(%)↑			
			2mm	2.5mm	3mm	4mm		2mm	2.5mm	3mm	4mm
3D UNet [30]	528.14	6.65 (3.57)	5.05	7.07	10.10	25.25	5.89 (3.34)	5.00	13.33	15.00	33.33
Multi-Phased Regression [15]	46.22	5.43 (2.67)	8.08	10.10	16.16	32.32	5.35 (2.35)	3.33	13.33	21.67	43.33
GeoRefineNet - HTC+Multiresolution	16.20	**2.29 (0.77)**	**48.82**	**67.34**	**77.11**	**89.56**	**3.89 (1.47)**	**27.86**	**43.39**	**56.73**	**65.30**

is markedly superior to that of other methods, including natural-domain landmark detectors like Hourglass and HRNet-W48, and medical-domain detectors like GU2Net and AFPP, which show significantly higher MREs and lower SDRs. Figure 3 presents the qualitative comparison of the proposed framework and suboptimal methods.

Table 2 presents the performance comparison of the proposed method with previous state-of-the-art methods on the XCAT CT and CQ500 datasets for the 3D cephalometric landmark detection task. For the XCAT CT dataset, the GeoRefineNet combined with HTC and Multiresolution learning demonstrates significant improvements over other methods. The MRE for GeoRefineNet - HTC and Multiresolution is 2.29 mm with a standard deviation of 0.77 mm, which is the lowest among all compared methods. The SDRs at 2 mm, 2.5 mm, 3 mm, and 4 mm are 48.82%, 67.34%, 77.11%, and 89.56%, respectively. These values indicate a substantial enhancement in detection accuracy, especially when compared to other methods such as 3D UNet and Multi-Phased Regression, which exhibit lower SDRs and higher MREs.

In the CQ500 dataset, the GeoRefineNet combined with HTC and Multiresolution learning also outperforms other methods. The MRE achieved is 3.89 mm with a standard deviation of 1.47 mm. The SDRs for 2 mm, 2.5 mm, 3 mm, and 4 mm are 27.86%, 43.39%, 56.73%, and 65.30%, respectively. This performance is markedly superior to that of other methods, including 3D UNet and Multi-Phased Regression, which show significantly higher MREs and lower SDRs. Furthermore, the XCAT CT (27 patients) and CQ500 datasets (22 patients) are small, rendering them susceptible to overfitting. However, the proposed framework demonstrated promising results on these small datasets and significantly outperformed the existing detection methods.

Ablation study: The ablation study results, as illustrated in Table 3, demonstrate the significant impact of the proposed framework components on the performance of the models on the XCAT CT and CQ500 datasets. For the XCAT CT dataset, FARNet without stages 2 and 3 shows an MRE of 2.91 mm and SDRs of 42.97%, 55.50%, 66.55%, and 80.90% under 2 mm, 2.5 mm, 3 mm, and 4 mm conditions, respectively. Incorporating stage 2 improves the MRE to 2.57 mm and SDRs to 46.38%, 60.19%, 71.15%, and 84.69%, while including both Stages 2 and 3, further enhances the MRE to 2.27 mm and SDRs to 53.47%, 67.57%, 77.48%, and 88.93%. Similarly, for HTC with Multiresolution learning, the absence of stage 3 results in an MRE of 1.93 mm and SDRs of

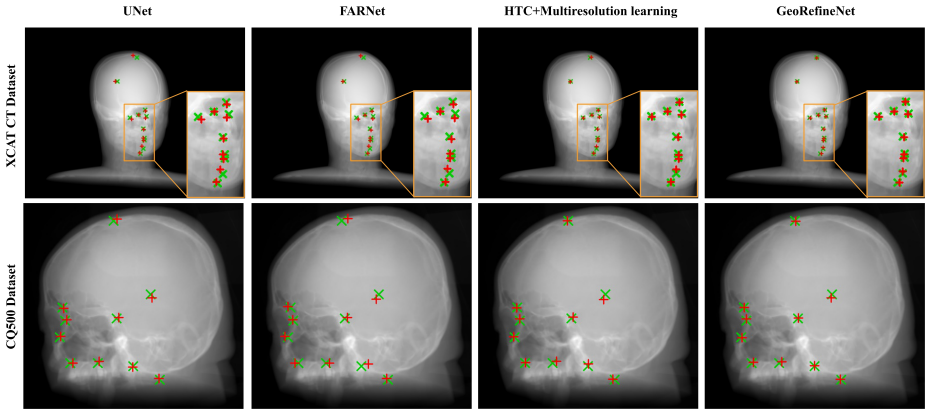

Fig. 3: Comparison of the proposed method and other models on the XCAT CT and CQ500 dataset in 2D landmark detection. Ground truth landmarks are green, and predictions are red.

Table 3: Comparison of the effects of the proposed framework components.

Model	Stage2	Stage3	XCAT CT dataset					CQ500 dataset				
			MRE(SD)↓	SDR(%)↑				MRE(SD)↓	SDR(%)↑			
				2mm	2.5mm	3mm	4mm		2mm	2.5mm	3mm	4mm
FARNet [2]	✗	✗	2.91 (3.44)	42.97	55.50	66.55	80.90	8.03 (13.97)	9.25	14.29	20.06	32.60
	✓	✗	2.57 (1.79)	46.38	60.19	71.15	84.24	7.37 (8.54)	11.04	18.19	24.01	37.41
	✓	✓	**2.27 (1.63)**	**53.47**	**67.57**	**77.48**	**88.93**	**6.23 (6.21)**	**13.78**	**20.72**	**26.83**	**40.97**
HTC+Multiresolution learning [24]	✗	✗	1.93 (2.27)	66.74	77.30	84.19	91.82	5.24 (5.39)	20.05	28.41	36.76	51.97
	✓	✗	1.75 (1.70)	67.75	79.58	85.94	92.76	4.93 (4.92)	22.62	32.25	42.35	56.64
	✓	✓	**1.61 (1.47)**	**75.76**	**84.62**	**89.72**	**95.06**	**4.58 (3.56)**	**25.61**	**35.84**	**45.38**	**60.39**

66.74%, 77.30%, 84.19%, and 91.82%, whereas the complete model with both stages achieves an MRE of 1.61 mm and SDRs of 75.76%, 84.62%, 89.72%, and 95.06%. On the CQ500 dataset, FARNet without stages 2 and 3 has an MRE of 8.03 mm and SDRs of 9.25%, 14.29%, 20.06%, and 32.60%. Adding stage 2 reduces the MRE to 7.37 mm and improves SDRs to 11.04%, 18.19%, 24.01%, and 37.41%. The full model with both stages achieves an MRE of 6.23 mm and SDRs of 13.78%, 20.72%, 26.83%, and 40.97%. For HTC with Multiresolution learning, excluding stage 3 results in an MRE of 5.24 mm and SDRs of 20.05%, 28.41%, 36.76%, and 51.97%, while the complete model yields an MRE of 4.58 mm and SDRs of 25.61%, 35.84%, 45.38%, and 60.39%. These results confirm that including stages 2 and 3 significantly enhances model performance, with the complete GeoRefineNet - HTC and Multiresolution configuration achieving the best accuracy and reliability in 2D cephalometric landmark detection.

4 Conclusion

The proposed method, GeoRefineNet, presents a novel multistage approach to localize cephalometric landmarks accurately in 2D and 3D CBCT scans. This

method improves on traditional techniques by applying information from multiple views and employing geometric insight to create 3D heatmaps, leading to more precise landmark detection. This method is more effective than the existing solutions, with better results on the XCAT and CQ500 datasets.

GeoRefineNet enhances landmark localization accuracy by using a 3D heatmap reconstruction and refinement process, combining landmark positions from multiple views to reduce the prediction errors of the 2D landmark detector. However, GeoRefineNet's reliance on geometric information from multiple views makes it unsuitable for applications that use only single-view images, such as cephalometric X-ray landmark detection. While specifically designed for cephalometric CBCT landmark detection, it can also be applied to other CBCT tasks, such as knee landmark detection.

This advancement has notable implications, especially for clinical practice in orthodontics and maxillofacial surgery. By enhancing the accuracy of landmark detection, GeoRefineNet supports more accurate diagnoses, better treatment planning, and improved patient outcomes. The ability of this method to address complex imaging data efficiently and reduce the chance of overfitting makes it a valuable tool for medical professionals. In the future, efforts will focus on bringing GeoRefineNet into everyday clinical use and testing its usefulness for other medical imaging tasks.

Acknowledgement

This work was partly supported by the Technology Development Program of MSS [S3146559], the National Research Foundation of Korea (NRF-2022R1-A2C1092072), and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT). (No. RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)).

References

1. Anwaar Khalid, M., Zulfiqar, K., Bashir, U., Shaheen, A., Iqbal, R., Rizwan, Z., Rizwan, G., Moazam Fraz, M.: Cepha29: Automatic cephalometric landmark detection challenge 2023. arXiv e-prints pp. arXiv–2212 (2022) 2
2. Ao, Y., Wu, H.: Feature aggregation and refinement network for 2d anatomical landmark detection. Journal of Digital Imaging pp. 1–15 (2022) 8, 10
3. 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) 8
4. Chen, R., Ma, Y., Chen, N., Liu, L., Cui, Z., Lin, Y., Wang, W.: Structure-aware long short-term memory network for 3d cephalometric landmark detection. IEEE Transactions on Medical Imaging **41**(7), 1791–1801 (2022) 2

5. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. The Lancet **392**(10162), 2388–2396 (2018) 3, 6
6. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020) 7
7. Dubey, S.R., Singh, S.K., Chaudhuri, B.B.: Activation functions in deep learning: A comprehensive survey and benchmark. Neurocomputing (2022) 6
8. Evans, C.A., Scarfe, W.C., Ahmad, M., Cevidanes, L.H., Ludlow, J.B., Palomo, J.M., Simmons, K.E., White, S.C.: Clinical recommendations regarding use of cone beam computed tomography in orthodontics. position statement by the american academy of oral and maxillofacial radiology. Oral Surgery Oral Medicine Oral Pathology Oral Radiology **116**(2), 238–257 (2013) 2
9. Feldkamp, L.A., Davis, L.C., Kress, J.W.: Practical cone-beam algorithm. Josa a **1**(6), 612–619 (1984) 5
10. Jiang, Y., Li, Y., Wang, X., Tao, Y., Lin, J., Lin, H.: Cephalformer: Incorporating global structure constraint into visual features for general cephalometric landmark detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III. pp. 227–237. Springer (2022) 2
11. Juneja, M., Garg, P., Kaur, R., Manocha, P., Batra, S., Singh, P., Singh, S., Jindal, P., et al.: A review on cephalometric landmark detection techniques. Biomedical Signal Processing and Control **66**, 102486 (2021) 7
12. Lang, Y., Lian, C., Xiao, D., Deng, H., Yuan, P., Gateno, J., Shen, S.G., Alfi, D.M., Yap, P.T., Xia, J.J., et al.: Automatic localization of landmarks in craniomaxillofacial cbct images using a local attention-based graph convolution network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. pp. 817–826. Springer (2020) 2
13. Miethke, R.: Possibilities and limitations of various cephalometric variables and analyses. Orthodontic Cephalometry. London: Mosby-Wolfe pp. 63–103 (1995) 2
14. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016) 8
15. Nishimoto, S., Saito, T., Ishise, H., Fujiwara, T., Kawai, K., Kakibuchi, M.: Three-dimensional craniofacial landmark detection in series of ct slices using multi-phased regression networks. Diagnostics **13**(11), 1930 (2023) 2, 9
16. Oh, K., Oh, I.S., Lee, D.W., et al.: Deep anatomical context feature learning for cephalometric landmark detection. IEEE Journal of Biomedical and Health Informatics **25**(3), 806–817 (2020) 2
17. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based cnns for landmark localization. Medical image analysis **54**, 207–219 (2019) 2
18. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 367–376 (2021) 8
19. Qian, J., Cheng, M., Tao, Y., Lin, J., Lin, H.: Cephanet: An improved faster r-cnn for cephalometric landmark detection. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 868–871. IEEE (2019) 2

20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 4, 8
21. Segars, W.P., Sturgeon, G., Mendonca, S., Grimes, J., Tsui, B.M.: 4d xcat phantom for multimodality imaging research. Medical physics **37**(9), 4902–4915 (2010) 6
22. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019) 8
23. Syben, C., Michen, M., Stimpel, B., Seitz, S., Ploner, S., Maier, A.K.: Technical note: Pyro-nn: Python reconstruction operators in neural networks. Medical Physics (2019) 7
24. Viriyasaranon, T., Ma, S., Choi, J.H.: Anatomical landmark detection using a multiresolution learning approach with a hybrid transformer-cnn model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 433–443. Springer (2023) 8, 10
25. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021) 8
26. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution vision transformer for dense predict. Advances in Neural Information Processing Systems **34**, 7281–7293 (2021) 8
27. Yun, H.S., Hyun, C.M., Baek, S.H., Lee, S.H., Seo, J.K.: Automated 3d cephalometric landmark identification using computerized tomography. arXiv preprint arXiv:2101.05205 (2020) 2
28. Zeng, M., Yan, Z., Liu, S., Zhou, Y., Qiu, L.: Cascaded convolutional networks for automatic cephalometric landmark detection. Medical Image Analysis **68**, 101904 (2021) 2
29. Zhang, K., Xu, N., Wu, J.: Multi-view fusion convolutional neural network for automatic landmark location on spinal x-rays. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–4. IEEE (2022) 2
30. Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D.: 3d deep learning for efficient and robust landmark detection in volumetric data. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. pp. 565–572. Springer (2015) 9
31. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: You only learn once: Universal anatomical landmark detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 85–95. Springer (2021) 8