ScaleCeph: A Two-Stage Coarse-to-Fine ROI Cascade for Cross-Device Cephalometric Landmark Localization

Ubaid Ur Rehman, Syed M. Adnan, Wakeel Ahmad, Huzaifa Ehsan, M. Nauman Tariq
Department of Computer Science, University of Engineering and Technology (UET), Taxila, Pakistan.
*Correspondence:24-MS-CS-01@uettaxila.students.edu.pk, ubaidurrehman721@gmail.com

Citation | Rehman. U. U, Adnan. S. M, Ahmad. W, Ehsan. H, Tariq. M. N, "ScaleCeph: A Two-Stage Coarse-to-Fine ROI Cascade for Cross-Device Cephalometric Landmark Localization", IJIST, Vol. 8, Issue. 5, pp 2081-2097, August 2026
Received | July 27,2026 **Revised |** Aug 15, 2026 **Accepted |** Aug 18, 2026 **Published |** Aug 20, 2026.

Cephalometric landmark localization is an important part of orthodontic diagnosis and treatment planning, but manual annotation is slow and has notable inter-observer variation. Automated methods have advanced rapidly, yet most are developed and tested on images from a single device and report a single aggregate error, leaving open how they behave when a clinic pools radiographs from several machines with different fields of view, contrast, and pixel spacing. We present ScaleCeph, a two-stage coarse-to-fine cascade that treats this cross-device variation as the central problem. A first HRNet-W32 localizes 19 landmarks on the whole radiograph; the bounding box of these predictions defines a craniofacial region of interest that is cropped and magnified to a canonical scale, absorbing the inter-device differences, and a second, independently trained network refines the landmarks on this magnified view. Coordinates are decoded with a robust windowed soft-argmax decoder, and ablations show that the design choices, prediction-driven crops, sharp $\sigma = 2$ heatmap targets with a weighted-BCE loss, and the windowed decoder, are each necessary rather than incidental. On the Aariz benchmark (1,000 cephalograms, seven devices), ScaleCeph attains a mean radial error (MRE) of 1.076 ± 1.18 mm (95% CI 0.99–1.16 mm) and 88.8% successful detection rate (SDR) within 2 mm (95% CI 87.7–90.0), against the single-stage baseline (1.252 ± 1.17 mm, 85.1%; paired Wilcoxon signed-rank $p < 10^{-20}$), and raises the macro-averaged per-device SDR@2 mm from 85.5% to 89.5%. On the best-resolved machines performance approaches inter-observer agreement, supporting prediction-driven scale normalization for robust cross-device localization.

Keywords: Cephalometric Landmark Localization; Automated Cephalometric Analysis; Cross-Device Robustness; Deep Learning; HRNet.

Introduction:

Cephalometric analysis is the identification of anatomical landmarks on lateral skull radiographs. It is vital to diagnosis, treatment planning, and growth assessment in orthodontics. [1][2]. Manual annotation is slow and shows large intra- and inter-observer variability. [3] which undermines diagnostic consistency. It led to automated cephalometric landmark detection becoming an active research area that focuses on accuracy, robustness, and clinical applicability. The digitization of radiographic workflows has driven extensive work on automated landmark detection with the adoption of artificial intelligence, specifically deep learning-based systems.

Deep learning has become the dominant paradigm in automated cephalometric landmark detection (CLD), with high-resolution feature representations especially prominent. High-Resolution Networks (HRNets) have gained notable attention. For instance, [4] tested HRNet with multiple resolutions on 2,000 private cephalograms (51 landmarks) and reached 1.08 mm MRE / 89.0% SDR@2 mm, while [5] proposed an HRNet with SRPose trained on 1,019 X-rays (22 landmarks), achieving 0.97 ± 0.52 mm MRE / 95.41% SDR@2 mm. [6] also proposed Self-CephaloNet (HRNetV2 + Self-ONN, coarse-to-fine), reporting 1.08 mm / 82.25% on the ISBI-2015 dataset (400 images, 19 landmarks) and 2.79 mm / 75.95% on PKU.

Coarse-to-fine and cascaded designs improve precision by refining an initial estimate through patch-based refinement or multi-resolution feature fusion. [7] used MLP refinement on HRNet predictions on 2,320 private images (23 landmarks) and achieved 0.42 ± 0.15 mm MRE for malocclusion Class II. [8] proposed a cascaded RetinaNet and U-Net approach for a 2,903-image private dataset (9 landmarks) and achieved 1.26 mm mean localization error with an 83.2% detection rate. [9] used a similar RetinaNet–U-Net pipeline (120 images; 18 landmarks) and showed 1.45 ± 0.92 mm MRE.

Furthermore, [10] used a VGG-16 stage that regresses the craniofacial region for scale invariance, and a second FPN stage with heatmap regression to localize the landmarks, reaching 1.789 ± 6.548 mm MRE and 78.44% SDR@2 mm on the Aariz dataset (29 landmarks). Additionally, [11] proposed a framework (CEPHMark-Net) using a single CNN backbone with multi-resolution semantic feature fusion for the coarse stage and a multi-head loss for refinement on the ISBI-2015 dataset (19 landmarks), achieving 1.23 ± 0.89 mm MRE and 87.17% SDR.

Ensemble and multimodal strategies have also been explored for robustness. [12] ensembled a network of UNet and UNet++, with an XGBoost combiner, which outperformed single models, reaching 1.27 mm / 83.6% on ISBI-2015 and 0.64 mm / 95.4% on a private set (700 images, 19 landmarks). While, [13] proposed GU2Net, a universal model trained jointly on head, hand, and chest landmarks with separable and dilated convolutions, achieving 1.54 ± 2.37 mm / 77.79% on the ISBI-2015 head data. [14] also combined a weighted ensemble of UNet, VNet, HRNet, and GU2Net, with a multi-scale cross-attention framework, which obtained 1.44 mm / 82.58% on ISBI-2015 and 1.73 mm / 77.95% on the MICCAI-2023 CL-Detection dataset (600 cephalograms, 38 landmarks).

In a parallel line, localization has also been reframed as object detection. [15] applied Mask R-CNN (ResNet-101–FPN with an RPN) on ISBI-2015, reaching a 98.29% overall detection rate but only 68.27% within 2 mm (2.19 mm mean error). In contrast, more recent efforts have utilized single-stage detectors as well; [16] used YOLOv5 for 17 landmarks on 350 private radiographs (0.95 mm / 90.44%), which is an adaptation of standard object detection models. Similarly, [17] used YOLOv8 with 295 layers on ISBI-2015 and reached an SDR of 86.31% under 2 mm.

Recently, some researchers have explored specialized and emerging architectures. [18] proposed CephRes-MHNet, a lightweight multi-head residual CNN with residual encoding, dual (spatial and channel) attention, and multi-head decoding, which reached 1.23 mm / 85.5%

on the Aariz dataset (19 landmarks). In contrast, [19] evaluated vision transformers with direct coordinate prediction against CNN (ConvNeXtV2) and heatmap-based transformer (SegFormer) baselines, which reached better generalization with 2.38mm MRE on a large private sparsely annotated dataset with 9 to 55 landmarks. Complementarily, [20] proposed a U-Net-based prototypical network followed by holistic prototype estimation with cross-image prototype alignment, masked relation mining, a multi-head self-attention layer, and positional embeddings, achieving 1.05 mm / 89.13% on a 1,000-image private set (10 landmarks). Table 1 summarizes the reviewed literature, highlighting the datasets, methodological approaches, evaluation protocols, and reported performance of existing automated cephalometric landmark localization methods.

Table 1. Summary of the reviewed literature on cephalometric landmark localization.

Ref No.	Pb yr	Dataset Details				Technique Used	Results (MRE, SDR@2 mm)
		Name	Img types	No. of Imgs	No. of LdMk		
[13]	'21	ISBI'15, ipila-hand, ChestXray	Ceph Xrays	400+ 909 + 279	19, 37, 6	GU2Net + Dilated Convolutions	1.54 ± 2.37 mm, 77.79%
[4]	'23	Private	Lateral Ceph Xrays	2300	51	High-Resolution Net (HRNet)	1.08 ± 0.87 mm, 89.00%
[15]	'24	ISBI 2015	Ceph Xrays	400	19	Mask R-CNN (ResNet-101-FPN + RPN)	SDR: 68.27%; Avg Error: 2.19 mm
[7]	'24	Private	Lateral Facial Photos	2300	23	HRNetV2 + MLP, heatmap regression	Class II: 0.42 ± 0.15 mm; Class III: 0.46 ± 0.16 mm
[16]	'24	Private	Ceph Xrays	350	17	YOLOv5	0.95 mm, 90.44%
[8]	'24	Private	Frontal Ceph Xrays	2,903	9	Cascaded CNNs: RetinaNet + U-Net	1.26 ± 1.94 mm, 83.2%
[9]	'24	Private	Lateral Ceph Xrays	120	18	Cascaded CNNs: RetinaNet + U-Net	1.45 ± 0.92 mm
[10]	'24	Aariz	Lateral Ceph Xrays	1000	29	Cascaded CNNs: VGG-16 + Feature Pyramid Net (FPN), heatmap regression	1.789 ± 6.548 mm, 78.44%, ROI IoU ~0.86
[11]	'24	ISBI'15	Lateral Ceph Xrays	400	19	Two-stage (CNN-Semantic fusion + multi-head refinement loss)	1.23 ± 0.89 mm, 87.17%
[12]	'24	ISBI'15 + Private	Ceph Xrays	400+ 700	19	Base Models (UNet, UNet++) with Ensemble (XGBoost, DBSCAN)	1.27 mm, 83.61%; Prvt: 0.69 mm, 95.4%

Ref No.	Pb yr	Dataset Details				Technique Used	Results (MRE, SDR@2 mm)
		Name	Img types	No. of Imgs	No. of LdMk		
[17]	'24	ISBI'15	Ceph Xrays	400	19	YOLOv8	SDR: 86.31%
[20]	'24	CephAdoAdu (private)	Ceph Xrays	1000	10	U-Net + Prototypical Network: holistic LdMk prototypes, alignment loss, self-attention	1.05 ± 0.33 mm, 89.13%
[14]	'25	ISBI'15, MICCAI 2023	Ceph Xrays	400+ 600	19, 38	Ensemble models: UNet, VNet, HRNet, GU2Net + cross-attention	ISBI: 1.44 mm, 82.58%; MICCAI: 1.73 mm, 77.95%
[6]	'25	ISBI'15	Ceph Xrays	400	19	Self-CephaloNet (HRNetV2 + Self-ONN)	1.08 mm, 82.25%
[5]	'25	Private	Lateral Ceph Xrays	1019	22	CNN-based (HRNet, SRPose, heatmap regression)	0.97 mm, 95.41%
[18]	'25	Aariz	Lateral Ceph Xrays	1,000	19	Multi-Head Residual CNN + heatmap regression	1.23 mm, 85.5%
[19]	'25	ISBI'15 + 2 PVT	Ceph Xrays	600+ 1408+ 2687	19, 37, 9–55	ViT, direct coordinate prediction	2.38 mm (small ViT), 2.52 mm (large ViT)

Despite this progress, these recent methods reveal a gap. The strongest reported results are obtained under fixed acquisition conditions, while reliability under varying acquisition conditions remains largely uncharacterized. Three properties of the prevailing evaluation regime account for this. First, the lowest reported errors are obtained on single-center, single-scanner collections; such results describe accuracy under fixed acquisition and do not establish generalizability to images acquired on other units. Second, where broader datasets have been used, performance is still summarized as one error pooled over the whole test set. A pooled figure cannot distinguish a method that is uniformly reliable from one whose average is carried by the devices that dominate the sample, so device-level behavior remains invisible even when the data are multi-source. Third, the strategies most often used to improve robustness carry costs of their own: ensembling and multimodal fusion increase architectural and computational complexity. Additionally, direct coordinate regression with transformers has been reported to generalize less effectively than CNN heatmap baselines. Thus, the remaining research gap is not raw accuracy on a fixed test set but systematically evaluating per-device robustness when cephalometric radiographs acquired from multiple machines are pooled. As illustrated in Figure 1, four representative machines from the seven-device Aariz dataset exhibit differences in field of view, image contrast, and pixel spacing, highlighting the acquisition heterogeneity that ScaleCeph is designed to address.

The remainder of this paper is organized as follows. The Materials and Methods section describes the Aariz benchmark, the preprocessing pipeline, the two stages of the cascade, the prediction-driven region-of-interest rule, and the training protocol. The Results and Discussion section reports the comparison with the state of the art, the per-device analysis,

the ablation studies, and the per-landmark and qualitative behavior, and then discusses the sources of residual error and the limitations of the method. The Conclusion summarizes the findings.

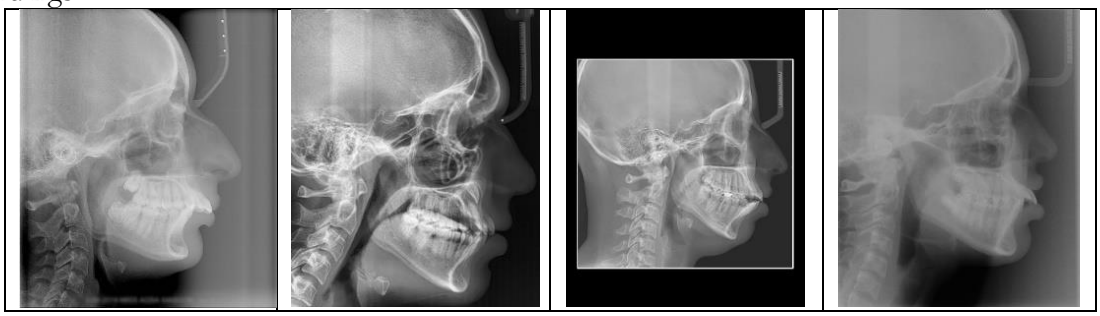

Figure 1. Sample cephalometric radiographs from four of seven different imaging machines showing variations in image contrast, brightness, and acquisition characteristics.

Research objectives:

The objective of this research is to obtain accurate cephalometric landmark localization that remains reliable across radiographs acquired on different X-ray units, and to measure that reliability per device rather than as a single pooled error.

Novelty statement:

The novelty of the work lies in treating the region-of-interest crop as a per-image device scale normalizer driven entirely by the first stage's own predictions, thereby addressing inter-device variation without using ground-truth landmarks to define the ROI.

Our contributions are: (1) a scale-normalizing ROI cascade designed to improve localization across the seven heterogeneous imaging devices; (2) the first full per-device analysis on the Aariz benchmark that makes cross-device robustness an explicit evaluation outcome rather than a single pooled number; (3) a prediction-driven cropping rule that uses the first stage's predictions, never the ground truth, to define the ROI in both training and inference, eliminating the train/inference mismatch in the cascade; and (4) an ablation-justified heatmap regime and a robust windowed soft-argmax decoder. Together, these components establish ScaleCeph as a coarse-to-fine framework specifically designed to improve robust cephalometric landmark localization across heterogeneous imaging devices.

Material and Methods:

Dataset:

We evaluate on the Aariz (CEPHA29) benchmark dataset [21], a publicly available collection of 1,000 lateral cephalograms acquired on seven different X-ray units and annotated with 29 landmarks; following the previous benchmark protocol ISBI-15 [22], we train and report on the 19-landmark subset and the fixed, subject-disjoint 700/150/150 train/validation/test split. The seven devices differ substantially in acquisition characteristics: field of view, contrast, and, most importantly, physical resolution, with isotropic pixel spacing ranging from 0.089 to 0.144 mm/px, a 1.6-fold spread (Table 2). They are also represented in very unequal proportions (e.g., 256 training images from ART Plus versus 29 from ProMax 2D), which motivates the machine-balanced sampling and macro-per-machine model selection described below. Ground-truth landmarks were established through the dataset's two-phase annotation protocol: two junior orthodontists marked them two senior orthodontists reviewed and corrected the markings. The ground truth for each landmark is the mean of the averaged junior and the averaged senior annotations. At the end of the review phase, the dataset reports an inter-observer variability of 0.425 ± 0.552 mm (mean radial error ± standard deviation) between the two senior orthodontists, which we use as a practical reference for the level of agreement achievable by human experts and as a target for automated methods. All errors are reported in millimeters using each device's pixel spacing.

Table 2. Acquisition devices in the Aariz dataset: isotropic pixel spacing and per-split image counts.

Device	Pixel spacing (mm/px)	Train	Val	Test
ART Plus	0.100	256	55	55
Hyperion X5	0.089	101	21	21
ProMax 2D	0.139	29	6	6
ProMax with ProTouch	0.139	95	20	20
Rotograph EVO	0.135	54	12	13
Smart3D	0.100	41	9	9
Veraviewepocs 2D	0.144	124	27	26
Total (7 devices)	**0.089 – 0.144**	**700**	**150**	**150**

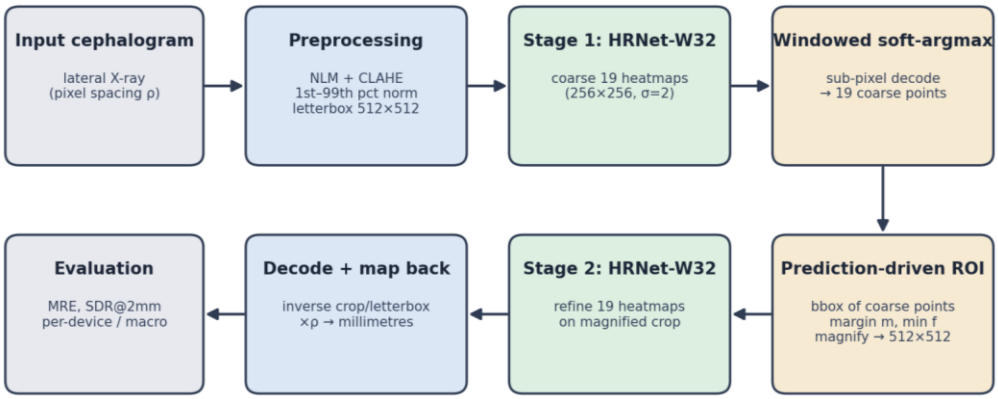

Figure 2. Complete ScaleCeph processing pipeline, from preprocessing to landmark prediction and millimeter-space evaluation.

Overview of ScaleCeph:

The end-to-end processing pipeline, from preprocessing to millimeter-space evaluation, is summarized in Figure 2. ScaleCeph is a two-stage coarse-to-fine cascade illustrated in Figure 3. Both stages are HRNet-W32 [23] heatmap regressors of identical architecture but trained independently; the crop between them normalizes device scale. Given a cephalogram, the task is to predict the image-plane coordinates of K = 19 landmarks and report error in millimeters using the device pixel spacing ρ.

Backbone selection:

HRNet-W32 is used in both stages because it maintains a high-resolution representation throughout the network while repeatedly fusing information across parallel resolution streams, a property well suited to precise sub-pixel localization and the reason HRNet has become a default backbone for landmark tasks [23]. It was deliberately preferred over recent transformer and hybrid alternatives: vision transformers with direct coordinate regression generalize less reliably than CNN heatmap models [19], because attention-based models are data-hungry and lack the built-in spatial inductive bias that heatmap CNNs exploit. Among HRNet widths, the W32 variant (≈31.4 M parameters per stage) was chosen as a balance between the representational capacity needed for sharp σ = 2 heatmaps and the training stability and memory cost of larger variants; ImageNet-pretrained initialization further shortens convergence on a dataset that is modest by deep-learning standards.

Image preprocessing:

Each radiograph is enhanced with non-local-means denoising followed by CLAHE, rescaled to [0,1] by its 1st–99th intensity percentiles, a per-image normalization that suppresses

the exposure differences between machines, and letterboxed to a 512×512 input while preserving aspect ratio.

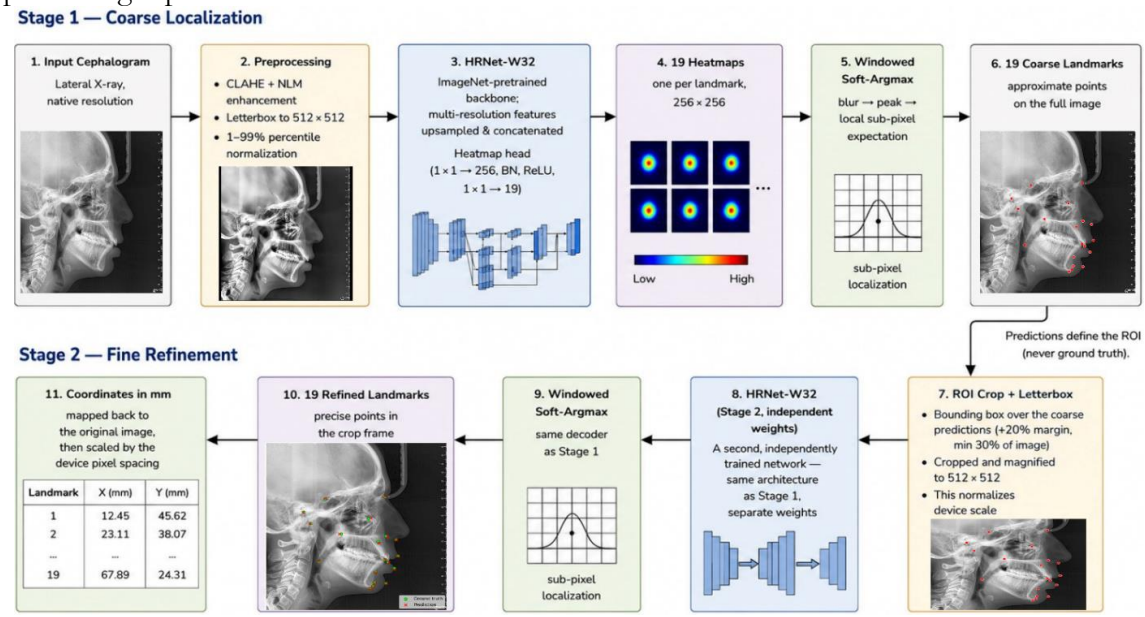

Figure 3. Overview of ScaleCeph multi-stage framework. Stage 1 predicts coarse landmarks on full images, leading to a prediction-driven ROI. Stage 2 refines positions within the ROI, which are mapped back and converted to millimeters using the device pixel spacing.

Stage 1: coarse localization:

An HRNet-W32 backbone (ImageNet-pretrained) yields multi-resolution features that are upsampled and concatenated, then mapped by a 1×1→256→BN→ReLU→1×1→19 head to 19 logit maps at 256×256. Targets are isotropic Gaussians drawn at heatmap resolution, as defined in Equation (1),

$$G_k(u) = \exp(-\| u - p_k \|^2 / (2\sigma^2)), \quad \sigma = 2 \quad (1)$$

where $G_k(\boldsymbol{u})$ is the target heatmap value at pixel $\boldsymbol{u}$ for landmark k (k = 1, …, 19); $\boldsymbol{u}$ = (u_x, u_y) is a pixel location on the 256×256 heatmap grid; $\boldsymbol{p}_k$ is the ground-truth position of landmark k expressed in heatmap coordinates; $\|\cdot\|$ denotes the Euclidean norm; and σ = 2 px is the Gaussian standard deviation, which sets the sharpness of the target.

Coordinates are decoded with a robust windowed soft-argmax, given in Equation (2). The per-pixel probability map is first smoothed with a Gaussian filter; the peak u* of the smoothed map is located by argmax; and a sub-pixel estimate is then formed as a temperature-weighted expectation over a small window W centered on that peak, evaluated on the unsmoothed map,

$$\hat{c}_k = \Sigma_{\{u \in W\}} \omega_u \cdot u,$$

$$u^* = \text{argmax}_u P_k(u), \quad \omega_u = \text{softmax}_{u \in W}\left(\frac{P_k(u)}{T}\right), \quad T = 0.05 \quad (2)$$

where ĉk is the decoded sub-pixel coordinate of landmark k; Pk is the per-pixel probability map for landmark k, obtained by applying a sigmoid to the predicted logits; P̃k is Pk after Gaussian smoothing with σ_{blur} = 2 px (9×9 kernel), which suppresses isolated high-response pixels before the peak is selected; u* is the location of the maximum of the smoothed map; W = {u: |ux − ux*| ≤ 6 and |uy − uy*| ≤ 6} is the 13×13 pixel window centred on u*; ωu is the weight assigned to pixel u, obtained as a softmax of Pk(u)/T restricted to W and therefore summing to one over W; and T = 0.05 is the temperature controlling the sharpness of the weighting.

Confining the expectation in Equation (2) to a window around a robustly chosen peak retains sub-pixel accuracy while avoiding the catastrophic outliers that a global argmax-Taylor (DARK) or global soft-argmax decoder produces on real heatmaps.

Region-of-interest extraction: The ROI is computed from the Stage-1 predictions, never from the ground truth, so that training and inference see the same distribution. Let (x_0, y_0) and (x_1, y_1) be the minimum and maximum coordinates of the 19 coarse predictions. The box is expanded by a margin m on each side and floored at a fraction f of the image dimensions, as defined in Equation (3),

$$w = \max\big((1+2m)(x_1 - x_0), fW_0\big), \mathrm{h} = \max\big((1 + 2\mathrm{m})(\mathrm{y}_1 - \mathrm{y}_0), \mathrm{f} \cdot \mathrm{H}_0\big), m = 0.20, f = 0.30 \quad (3)$$

where w and h are the width and height, in pixels, of the region-of-interest (ROI) box; x_0 and x_1 are the minimum and maximum predicted x-coordinates, and y_0 and y_1 are defined analogously for the y-axis. W_0 and H_0 are the original width and height in pixels. m is the fractional margin added on each side of the predicted landmark extent, and f is the lower bound on the box size, expressed as a fraction of the corresponding image dimension.

The box produced by Equation (3) is centered on the midpoint of the predicted landmark extent and clipped to the image boundaries. During training only, the box is jittered to make Stage 2 tolerant of coarse-stage error: its center is shifted by up to ±5% of the box size, and its width and height are each scaled by up to ±10%, applied independently per axis. No jitter is applied during validation or testing, where the box is deterministic.

Stage 2: fine localization and coordinate recovery: The ROI is cropped from the enhanced full-resolution image and letterboxed to 512×512, magnifying the skull to a canonical scale so that each structure is represented by more pixels and the inter-device field-of-view and resolution differences are largely absorbed before the network sees the image. A second, independently trained HRNet-W32 with the same head, targets (Equation (1)), and decoder (Equation (2)) refines the landmarks. Refined coordinates are mapped back through the crop and letterbox transforms to the original image plane and scaled by the device pixel spacing ϱ to millimeters.

Table 3. Implementation settings (identical for both stages).

Setting	Value	Setting	Value
Input / heatmap	512 / 256 (stride 2)	Optimizer	AdamW
Heatmap target σ	2 px	LR (backbone/head)	1e-4 / 2e-4
Loss / pos-weight	wBCE / 1500	Schedule	cosine + 5-ep warmup
Loss weights	1.0 / 0.1	Epochs / batch	150 / 4×4 (eff. 16)
Decoder	windowed soft-argmax	EMA / selection	0.999 / macro-per-machine
ROI margin / min	0.20 / 0.30	Params (cascade)	62.8 M

Training: Both stages are initialized from ImageNet-pretrained HRNet-W32 weights, with no frozen layers; Stage 2 is initialized independently rather than fine-tuned from Stage 1 and the two stages are trained separately. Both minimize a heatmap-dominant objective consisting of a weighted binary cross-entropy on the heatmap logits and a low-weight Smooth-L1 coordinate regression term, as expressed in Equation (4),

$$L = L_{hm} + \lambda_{ref} \cdot L_{ref}, \; L_{hm} = \mathrm{wBCE}(\mathrm{Z}, \mathrm{G}; w_+), w_+ = 1500), \lambda_{ref} = 0.1 \quad (4)$$

where L is the total training objective minimized by each stage; L_{hm} is the weighted binary cross-entropy between the predicted heatmap logits Z and the Gaussian targets G of Equation (1); $w_+ = 1500$ is the positive-class weight applied to the foreground term, which counters the foreground-to-background imbalance of a $\sigma = 2$ target on the 256×256 grid; L_{ref} is a Smooth-L1 loss ($\beta = 0.01$) between the predicted and ground-truth landmark coordinates,

both normalized to [0, 1]; and $\lambda_{ref} = 0.1$ is the weight on the coordinate term, which acts as an auxiliary constraint rather than the primary supervisory signal.

The positive weight counters the extreme foreground/background imbalance of a $\sigma = 2$ target; the ablation shows this regime is required (a plain MSE loss collapses). Networks are trained for 150 epochs with AdamW (discriminative learning rates), a cosine schedule, EMA, a machine-balanced sampler, and checkpoint selection by macro-per-machine validation error so that minority devices are not masked (Table 3).

Computational complexity and efficiency: Because the framework is intended for routine clinical use, we characterize its cost across both stages and both phases rather than reporting parameters alone. The cascade comprises two independent HRNet-W32 networks of 31.4 M parameters each (62.8 M total), operating on a 512×512 input and a 256×256 heatmap. Table 4 summarizes the theoretical cost (multiply-accumulate operations per stage) together with the measured training and inference footprint on a single commodity NVIDIA T4 GPU (16 GB). At inference the two forward passes total ≈100 ms/image (≈50 ms/stage); training used 150 epochs per stage under automatic mixed precision. Two independent passes make the model accurate rather than lightweight, but the whole pipeline trains and runs on a single 16 GB GPU without specialized hardware, which is the relevant practicality criterion for a chairside or clinic-server deployment.

Table 4. Computational complexity and efficiency of ScaleCeph. Memory is reported per stage, the cascade's peak memory equals a single stage's, not their sum.

Aspect	Stage 1	Stage 2	Full cascade
Parameters	31.4 M	31.4 M	62.8 M
MACs @512² (G)	76.2	76.2	152.3
Peak GPU memory – training	6.1 GB	6.1 GB	6.1 GB
Peak GPU memory – inference	—	—	1.1 GB
Training time	6.0 h	5.9 h	11.9 h
Inference latency	49.7 ms	49.7 ms	99.4 ms / image
Throughput	—	—	10.1 img/s

Results:

We evaluate on the Aariz benchmark: 1,000 lateral cephalograms from seven devices, 19 landmarks, split 700/150/150. All results are on the held-out 150-image test set; the inter-observer variability reported for this dataset is 0.425 ± 0.552 mm. The cascade has ≈62.8 M parameters and runs in ≈100 ms/image on one GPU.

Overall test performance: Table 5 compares ScaleCeph with methods evaluated on the Aariz dataset with standard ISBI-15 landmarks. ScaleCeph achieves the lowest MRE, 1.076 mm, and the highest SDR at every evaluated tolerance: 88.8% at 2 mm, 93.4% at 2.5 mm, 95.4% at 3 mm, and 98.0% at 4 mm. Compared with CephRes-MHNet [18], ScaleCeph reduces MRE by 0.154 mm and improves SDR@2 mm by 3.3 percentage points, and the gain holds at both stricter and broader tolerances. All methods in Table 5 are compared under an identical protocol; the same fixed, subject-disjoint 700/150/150 Aariz split and the same 19-landmark subset, with baseline figures taken from the same-split benchmark study [18]. ScaleCeph also clearly exceeds AFPF-Net, Revisit-HPE-SRHead, and the dataset cascaded-CNN baseline.

Table 5. Test performance on the standard Aariz-19 split.

Method	MRE (mm)	SDR@2	@2.5	@3	@4
Cascaded CNN [10]	1.789	78.4	85.5	89.5	94.4
Revisit-HPE-SRHead [24]	1.35	83.6	88.1	90.8	94.4
AFPF-Net [25]	1.25	84.1	88.9	92.0	95.8
CephRes-MHNet [18]	1.23	85.5	89.4	92.2	95.3

ScaleCeph (ours)	**1.076**	**88.8**	**93.4**	**95.4**	**98.0**

Cross-device performance: Table 6 and Figure 4 report the per-device comparison, which is the central result of this work. The cascade lowers MRE on all seven machines and raises SDR@2 mm on six, while maintaining the already high 99.3% on Hyperion X5; the macro average rises from 85.5% to 89.5% (1.21 → 1.01 mm), and the hardest device, Rotograph EVO, improves from 60.3% to 64.0%. These results support the interpretation that the prediction-driven ROI crop acts as a device-scale normalizer: because the comparison uses the same decoder and differs primarily in the presence of the refinement cascade, the consistent improvement across devices is consistent with the proposed normalization mechanism rather than with added capacity alone. The magnitude of the per-device estimates should nevertheless be read with caution where test sets are small, particularly ProMax 2D (n = 6) and Smart3D (n = 9), whose estimates carry greater sampling uncertainty.

Table 6. Per-device MRE / SDR@2 mm, single stage (S1) versus cascade (Casc).

Machine	N	S1 MRE ± SD	Casc MRE± SD	Casc 95% CI	S1 @2	Casc @2
ART Plus	55	1.46 ± 1.28	1.33 ± 1.38	1.25–1.44	81.3	85.1
Hyperion X5	21	0.50 ± 0.35	0.46 ± 0.35	0.41–0.54	99.3	99.3
ProMax 2D	6	1.27 ± 0.67	0.94 ± 0.56	0.84–1.03	86.8	94.7
ProMax ProTouch	20	1.59 ± 1.38	1.21 ± 1.34	1.06–1.42	77.6	86.6
Rotograph EVO	13	2.15 ± 1.50	1.95 ± 1.50	1.77–2.15	60.3	64.0
Smart3D	9	0.59 ± 0.39	0.47 ± 0.37	0.40–0.53	98.8	99.4
Veraviewepocs 2D	26	0.94 ± 0.58	0.73 ± 0.53	0.68–0.77	94.5	97.4
Macro	**—**	**1.21**	**1.01**	**—**	**85.5**	**89.5**

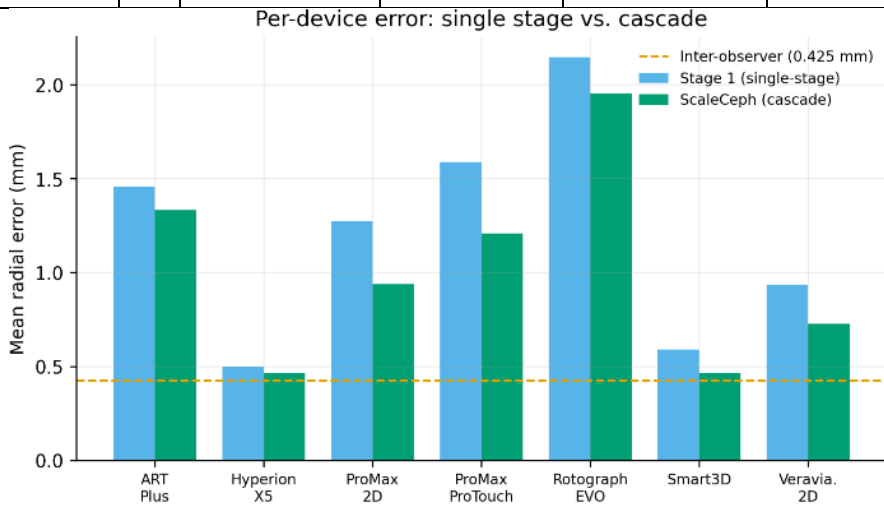

Figure 4. Per-device MRE for the single-stage baseline and the ScaleCeph cascade. The dashed line marks the 0.425 mm inter-observer agreement reported for the Aariz dataset.

Ablation studies:

The ablations in Table 7 justify each design choice. The cascade is the largest effect, cutting MRE from 1.252 to 1.076 mm (a 14.1% relative reduction) and raising SDR@2 mm from 85.1% to 88.8%. A paired Wilcoxon signed-rank test on per-image error confirms the improvement is significant ($p < 10^{-20}$), with 18 of 19 landmarks improved (16 of 19 surviving Bonferroni correction, $\alpha = 0.0026$). The heatmap regime is decisive: under the sharp $\sigma = 2$ target a plain MSE loss (30.6%) and an Adaptive-Wing loss (66.5%) fail to converge on the shared learning-rate schedule, whereas weighted BCE reaches 84.7%. Weighted BCE gives a lower error on every one of the 150 test images (Wilcoxon $W = 0$, $p \approx 2 \times 10^{-26}$, the minimum attainable at n = 150). This comparison is read as an effect size (1.25 mm versus 34.2 and 26.9

mm). For the decoder, the windowed and global soft-argmax variants are not separable on mean accuracy, the global decoder is in fact lower (1.065 vs 1.247 mm) and higher on SDR@2 mm (88.0 vs 84.7%). They differ in the extreme tail, a signed-rank test on the per-image worst-landmark error favors the windowed decoder (3.50 vs 3.87 mm, $p = 0.028$). Additionally, it bounds single largest error at 19.3 mm whereas the global decoder gives 74.9 mm error, so we retain the windowed decoder for its bounded worst case rather than for higher mean accuracy. The windowed decoder also matches DARK in the mean (84.7 vs 84.8%), and CLAHE is overall neutral yet lifts the low-contrast Rotograph device by about 2 points.

Table 7. Single-variable ablations. The chosen configuration is in bold.

Study	Configuration	MRE	SDR@2
Pipeline	Stage 1 only	1.252	85.1
	+ ROI cascade	**1.076**	**88.8**
Loss	MSE / Adaptive-Wing	34.2 / 26.9	30.6 / 66.5
	Weighted BCE	**1.247**	**84.7**
Decoder	DARK / soft-argmax	1.259 / 1.065	84.8 / 88.0*
	Windowed soft-argmax	**1.247**	**84.7**
Preproc.	without / with CLAHE	1.253 / 1.247	84.9 / 84.7

* Global soft-argmax attains a higher SDR@2 mm but a heavy outlier tail (SD 1.80 mm, max 74.9 mm).

Hyperparameter sensitivity:

We assess sensitivity at inference time on the trained cascade, varying each constant that ScaleCeph introduces while holding the others fixed and evaluating on the same 150-image test set (Table 8). Across the ROI margin m, the minimum-ROI fraction f and the decoder temperature T, SDR@2 mm spans 87.96–89.40%, a range of 1.44 percentage points. The decoder temperature is effectively inert, over a five-fold range it changes SDR@2 mm by 0.11 points and MRE by 0.002 mm. The minimum-ROI fraction f is a safeguard against a degenerate coarse prediction; on this test set the coarse bounding box never fell below the floor, so the constraint was never active and the three rows are identical, a null result confirming that the safeguard was not triggered, rather than evidence that the box size is unimportant. The margin m is the only parameter with a systematic effect: a tighter crop raises the effective resolution of the Stage-2 input, and SDR@2 mm falls from 89.40% to 88.81% to 87.96% as m increases from 0.10 to 0.30 (MRE 1.062, 1.076, and 1.100 mm, respectively). At $m = 0.10$, however, one ground-truth landmark on one radiograph falls outside the cropped region and cannot be recovered, whereas $m = 0.20$ retains every landmark on every test image with at least 64 px of margin. Because Stage 2 was trained on crops generated with $m = 0.20$, the settings at either end of the sweep also introduce a mild train–inference distribution shift, so the trend in m should be read as the combined effect of resolution and crop-distribution match. We therefore fix $m = 0.20$ as a coverage guarantee rather than selecting it for test accuracy, the same bounded-worst-case reasoning that motivates the windowed decoder. The heatmap standard deviation σ and the learning rates are training-time constants carried over unchanged from the single-stage configuration (Table 3); σ is examined in the ablation study rather than in this inference-time sweep.

Table 8. Sensitivity of test performance to key hyperparameters (one factor varied at a time).

Hyperparameter	Value	MRE (mm)	SDR@2
ROI margin m	0.10	1.062	89.40
	0.20 (default)	**1.076**	**88.81**
	0.30	1.100	87.96
Min ROI fraction f	0.20	1.076	88.81
	0.30 (default)	**1.076**	**88.81**

	0.40	1.076	88.81
Decoder temp. T	0.02	1.074	88.77
	0.05 (default)	**1.076**	**88.81**
	0.10	1.076	88.70

Per-landmark and qualitative analysis:

Figure 5 shows best-case and worst-casepredictions. The error distribution is tight with 0.79 mm median and 2.12 mm in the 90th percentile, and 13 of the 19 landmarks fall below 1 mm. Table 9 reports the mean radial error and the 2 mm detection rate for each of the 19 landmarks. The overall distribution is mostly concentrated (median 0.79 mm, 90th percentile 2.12 mm, maximum 21.7 mm), and per-landmark errors are given in Table 9. The closest to the inter-observer floor are Sella (0.63 mm), Pogonion (0.71 mm), Menton (0.76 mm), Gnathion (0.80 mm) as our best-defined points. Residual error concentrates on the posterior, low-contrast landmarks like Articulare (1.67 mm), Porion (1.65 mm), Gonion (1.55 mm), and PNS (1.44 mm), where even experts' annotation is variable due to overlapping structures and weak edges. These same posterior landmarks turn out to be the hardest ones on the lowest-contrast device and appear to be limited by the imaging geometry rather than by the model. Articulare, Porion, and Gonion lie where bilateral bony structures overlap in a 2-D lateral projection, and PNS lies in a low-contrast region, so their position is ambiguous in the radiograph and expert annotators disagree there. Scale normalization magnifies these regions but cannot recover an edge that the acquisition never captured, which is a plausible explanation for why they remain the dominant contributors to residual error after refinement.

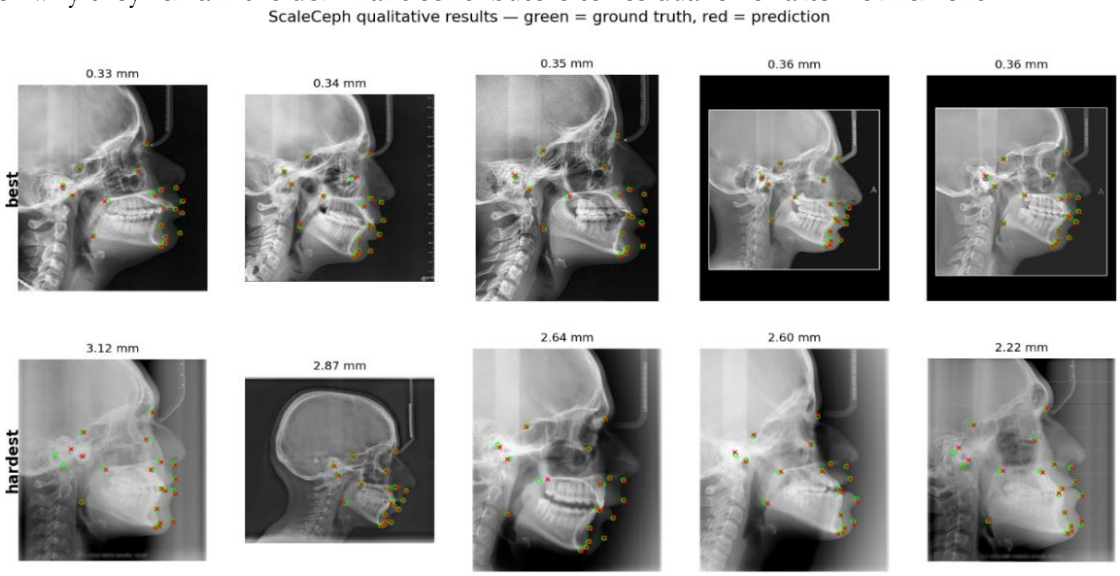

Figure 5. ScaleCeph qualitative results with per-image MRE above. Ground truth is green; Predictions are red markers. Top row shows best-performing images and bottom shows hardest images.

Table 9. Per-landmark MRE and 2 mm SDR on the Aariz-19 test set. The four hardest points (Ar, Po, Go, PNS) are posterior or low-contrast landmarks.

Landmark	MRE (mm)	SDR@2	Landmark	MRE (mm)	SDR@2
Sella (S)	0.63	99.3	Gonion (Go)	1.55	72.7
Nasion (N)	0.83	94.7	Lower Incisor Tip	0.84	92.7
Orbitale (Or)	1.22	80.0	Upper Incisor Tip	0.78	94.7
Porion (Po)	1.65	76.7	Labrale sup. (Ls)	0.94	96.7
Upper Inc. Apex	1.17	84.0	Labrale inf. (Li)	1.18	84.7
B-point (B)	0.89	95.3	Subnasale (Sn)	0.82	96.7
Pogonion (Pog)	0.71	97.3	Soft-t. Pog	1.26	86.7

Menton (Me)	0.76	99.3	PNS	1.44	80.7
Gnathion (Gn)	0.80	98.7	ANS	1.30	79.3
			Articulare (Ar)	1.67	77.3
Average	**1.076**	**88.8**	-	-	-

Discussion:

Why prediction-driven scale normalization works:

Our framework's (ScaleCeph) target claim is that what appears to be cross-device difficulty in cephalometric landmark localization is mechanically a scale problem that can be removed geometrically rather than learned around. If the improvement were concentrated on only one or two devices, it could plausibly reflect device-specific adaptation or simply the additional capacity of the second stage. Instead, the improvement seems to occur across all seven devices while the backbone, loss, and decoder remain fixed. This consistent pattern suggests that that the ROI is a scale-normalizing operation rather than merely an additional prediction stage. The magnitude also tracks the amount of normalization available in devices like Hyperion X5, which is the finest-sampled unit at 0.089 mm/px, and has the least scale variations and correspondingly gains the least.

Why the crop must be prediction-driven:

The closest methodological neighbor to this work is the two-stage cascaded CNN of [10], which also isolates a craniofacial region before refinement on the same dataset. It differs in one decisive respect: the region is produced by a dedicated VGG-16 stage trained to regress a box whose ground truth is derived from the annotated landmarks. Training the refinement stage on such ground-truth-derived crops can introduce a potential train-inference distribution mismatch, since at deployment the refinement stage sees only predicted regions. ScaleCeph removes the ROI regressor entirely and derives the box from the Stage-1 predictions themselves, so both stages see the same distribution in training and inference. On the same benchmark the result is 1.076 mm versus 1.789 mm and 88.8% versus 78.44% SDR@2 mm. We do not attribute the whole gap to the cropping rule, since the backbones also differ, but the comparison shows that a two-stage ROI cascade is not automatically effective; how the region is obtained matters.

Residual errors and what the method cannot resolve:

Error is not distributed evenly. It concentrates on Rotograph EVO, the lowest-contrast device, which remains at 1.95 mm and 64.0% SDR@2 mm after refinement, and on the posterior, low-contrast landmarks. Thirteen of nineteen landmarks fall below 1 mm. This pattern is consistent throughout the literature and is informative about what scale normalization can: magnifying the region of interest supplies resolution, but it cannot supply an edge that the acquisition did not record. Where the anatomical boundary is genuinely ambiguous in the radiograph, human annotators disagree as well, and the remaining error is at least partly irreducible with respect to the available ground truth. Improving these landmarks plausibly requires contrast-targeted enhancement or explicit anatomical constraints rather than further geometric normalization.

Decoder robustness:

The decoder ablation records a deliberate trade. A globally trained soft-argmax attains a higher SDR@2 mm than the windowed decoder (88.0% against 84.7% at the single-stage baseline) and a lower mean error, and its error percentiles up to the 99th are comparable to the windowed decoder's. The two are separable only in the extreme tail: the global decoder's single worst error reaches 74.9 mm. It's a Gonion failure in which the global expectation is dragged onto a secondary mode against a bounded 19.3 mm for the windowed decoder, and a signed-rank test on the per-image worst-landmark error favors the windowed decoder ($p = 0.028$). We therefore selected the windowed decoder on robustness grounds despite the lower aggregate score. A higher average SDR does not necessarily mean better clinical robustness if

the decoder occasionally produces catastrophic failures. So, bounded worst case is the safer default for clinical deployment, while reporting the global decoder's superior mean accuracy explicitly.

Clinical implications:

Cephalometric measurements such as the ANB angle, and the skeletal classifications derived from them, are computed directly from landmark coordinates, so a localization error that varies with the acquiring machine propagates into device-dependent measurements. The 2 mm threshold used throughout this work corresponds to the tolerance conventionally regarded as clinically acceptable for cephalometric landmark identification. By normalizing scale on a per-image basis, ScaleCeph achieves a macro-averaged SDR@2 mm of 89.5% across the seven units, with error reduced on every device. This is relevant to the common setting in which a practice pools cephalograms from several X-ray machines or receives referrals imaged elsewhere, as it indicates that accuracy does not depend strongly on the source device; whether this reduces the manual re-tracing clinicians currently perform on automated output remains to be measured.

On the best-resolved machines the mean radial error approaches the inter-observer agreement of 0.425 mm reported for this dataset, so the residual error is of the same order as the disagreement between expert annotators. Equally relevant to safe deployment, the per-device analysis makes reliability legible: it identifies the units on which predictions are least accurate, here Rotograph EVO, and on which expert review is therefore most warranted, rather than concealing that variation within a single pooled score. These properties are consistent with use as a decision-support tool for orthodontic diagnosis and treatment planning across heterogeneous imaging environments, subject to the validation on derived cephalometric measurements set out below. Final clinical responsibility remains with the practitioner.

Limitations:

Some limitations bound these results. The method is evaluated on a single benchmark and split, so cross-dataset behavior is untested. Two devices contribute to only 6 and 9 test images, so their per-device statistics widen confidence intervals and should be read as indicative rather than definitive. The reported figures derive from a single training run per stage, so run-to-run variance is not characterized. The cascade requires two forward passes at approximately 100 ms per image on a single NVIDIA T4, which is acceptable for routine clinical use but heavier than single-stage alternatives. No device-aware calibration is performed. Finally, the evaluation measures geometric error only, not the cephalometric angles and skeletal classifications that also drive treatment decisions.

Recommendations and future directions:

Five directions follow from the residual-error and cost analyses above. Firstly, *anatomy-aware refinement:* because the residual error is anatomically structured on the posterior, overlapping landmarks (Ar, Po, Go, PNS), an explicit anatomy-aware or graph-based refinement module is a promising route to further accuracy. Such a module could allow a confidently located landmark to constrain an uncertain neighbor, complemented by higher-resolution or point-specific crops for these landmarksto constrain an uncertain neighbor is a promising route to further accuracy, complemented by higher-resolution or point-specific crops for these landmarks. Secondly, a shared or partially shared backbone, or a lighter second stage, should be investigated to retain cascade accuracy at fewer parameters and lower latency to further reduce inference cost. Thirdly, robustness should be tested beyond a single benchmark through cross-dataset transfer, for example to ISBI-2015, and, most directly, through a leave-one-machine-out protocol that measures accuracy on a genuinely unseen device. Fourthly, for Device-aware calibration, Conditioning the model on, or adapting it to, the acquiring unit is a natural route to narrowing the gap on the hardest machines.

Finally,Validation on derived cephalometric measurements and skeletal classifications, against expert readings, should precede clinical deployment.

Conclusion:

This work addressed cross-device robustness in cephalometric landmark localization, the realistic setting in which a clinic pools radiographs from several X-ray units that differ in field of view, contrast, and resolution. We proposed ScaleCeph, a two-stage coarse-to-fine cascade in which a prediction-driven region-of-interest crop, derived from the first stage's own predictions rather than from the ground truth, magnifies the craniofacial region to a canonical scale and so normalizes device-to-device differences before a second network refines the landmarks. On the Aariz split, ScaleCeph attained a mean radial error of 1.076 mm and a successful detection rate within 2 mm of 88.8%, surpassing the prior state of the art on the same split, and the per-device analysis showed improved accuracy on all seven machines, raising the macro-averaged detection rate from 85.5% for the single-stage baseline to 89.5%. The principal contribution is a scale-normalizing cascade that makes cross-device robustness a measured, per-machine outcome rather than an assumed property. Its main costs are the two-pass, 62.8 M-parameter design and a residual error concentrated on the lowest-contrast device and the posterior landmarks; the directions set out above indicate how each might be addressed.

Acknowledgement.: The authors thank UET Taxila, for providing quality education, and acknowledge the creators of the Aariz dataset for making the dataset publicly available.

Conflict of interest: The authors declare no conflict of interest.

Project details: No specific grants from any funding agency were received.

References:

[1] "Does orthodontic treatment affect patients' quality of life? - PubMed." Accessed: Aug. 10, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18676797/

[2] J. Monisha, U. Sangeetha, B. Nivethitha, and B. Madhan, "Agreement between cephalometric analyses in diagnosing the dento-skeletal characteristics of malocclusion," *J. Oral Biol. Craniofacial Res.*, vol. 15, no. 4, pp. 744–748, Jul. 2025, doi: 10.1016/j.jobcr.2025.04.012.

[3] H. Zhang *et al.*, "Deep Learning Techniques for Automatic Lateral X-ray Cephalometric Landmark Detection: Is the Problem Solved?," Sep. 2024, Accessed: Jul. 27, 2026. [Online]. Available: https://arxiv.org/pdf/2409.15834

[4] Q. Chang, Z. Wang, F. Wang, J. Dou, Y. Zhang, and Y. Bai, "Automatic analysis of lateral cephalograms based on high-resolution net," *Am. J. Orthod. Dentofac. Orthop.*, vol. 163, no. 4, pp. 501-508.e4, Apr. 2023, doi: 10.1016/J.AJODO.2022.02.020.

[5] M. Han *et al.*, "Automated Landmark Detection and Lip Thickness Classification Using a Convolutional Neural Network in Lateral Cephalometric Radiographs," *Diagnostics 2025, Vol. 15, Page 1468*, vol. 15, no. 12, p. 1468, Jun. 2025, doi: 10.3390/DIAGNOSTICS15121468.

[6] M. S. I. Sumon *et al.*, "Self-CephaloNet: a two-stage novel framework using operational neural network for cephalometric analysis," *Neural Comput. Appl. 2025 3716*, vol. 37, no. 16, pp. 9777–9805, Mar. 2025, doi: 10.1007/S00521-025-11097-6.

[7] Y. Shimamura *et al.*, "Accuracy of cephalometric landmark and cephalometric analysis from lateral facial photograph by using CNN-based algorithm," *Sci. Reports 2024 141*, vol. 14, no. 1, pp. 31089-, Dec. 2024, doi: 10.1038/s41598-024-82230-z.

[8] S. H. Han *et al.*, "Accuracy of posteroanterior cephalogram landmarks and measurements identification using a cascaded convolutional neural network algorithm: A multicenter study," *Korean J. Orthod.*, vol. 54, no. 1, pp. 48–58, 2024, doi: 10.4041/KJOD23.075.

[9] S. Kang, I. Kim, Y. J. Kim, N. Kim, S. H. Baek, and S. J. Sung, "Accuracy and clinical

validity of automated cephalometric analysis using convolutional neural networks," *Orthod. Craniofac. Res.*, vol. 27, no. 1, pp. 64–77, Feb. 2024, doi: 10.1111/OCR.12683.

[10] R. Khan, M. A. Khalid, K. Zulfiqar, U. Bashir, and M. M. Fraz, "Enhancing Cephalometric Landmark Detection with a Two-Stage Cascaded CNN on Multi-resolution Multi-modal Data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14860 LNCS, pp. 3–18, 2024, doi: 10.1007/978-3-031-66958-3_1/SAVE-RESEARCH.

[11] M. A. Khalid, A. Khurshid, K. Zulfiqar, U. Bashir, and M. M. Fraz, "A two-stage regression framework for automated cephalometric landmark detection incorporating semantically fused anatomical features and multi-head refinement loss," *Expert Syst. Appl.*, vol. 255, p. 124840, Dec. 2024, doi: 10.1016/J.ESWA.2024.124840.

[12] S. Rashmi, S. Srinath, R. Rakshitha, and B. V. Poornima, "Ensemble learning methods with single and multi-model deep learning approaches for cephalometric landmark annotation," *Discov. Artif. Intell. 2024 41*, vol. 4, no. 1, pp. 93-, Nov. 2024, doi: 10.1007/S44163-024-00207-3.

[13] H. Zhu, Q. Yao, L. Xiao, and S. K. Zhou, "You Only Learn Once: Universal Anatomical Landmark Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12905 LNCS, pp. 85–95, Mar. 2022, doi: 10.1007/978-3-030-87240-3_9.

[14] Z. Zhu *et al.*, "An ensemble-based deep learning method through multi-scale cross-attention training for cephalometric landmark localization on lateral X-ray images," Mar. 2025, doi: 10.21203/RS.3.RS-6105085/V1.

[15] Z. Jiao *et al.*, "Deep learning for automatic detection of cephalometric landmarks on lateral cephalometric radiographs using the Mask Region-based Convolutional Neural Network: a pilot study," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.*, vol. 137, no. 5, pp. 554–562, May 2024, doi: 10.1016/j.oooo.2024.02.003.

[16] S. A. Prativi, A. B. Suksmono, T. L. E. Rajab, D. Danudirdjo, and A. Laviana, "Automatic landmark detection in cephalometric lateral radiograph using deep learning one-stage detectors," *2024 14th Int. Conf. Syst. Eng. Technol. ICSET 2024 - Proceeding*, pp. 67–72, 2024, doi: 10.1109/ICSET63729.2024.10775273.

[17] I. Tafala, F. E. Ben-Bouazza, A. Edder, O. Manchadi, M. Et-Taoussi, and B. Jioudi, "Cephalometric Landmarks Identification Through an Object Detection-based Deep Learning Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 859–867, 2024, doi: 10.14569/IJACSA.2024.0150286.

[18] A. Jaheen *et al.*, "CephRes-MHNet: A Multi-Head Residual Network for Accurate and Robust Cephalometric Landmark Detection," Nov. 2025, Accessed: Jul. 27, 2026. [Online]. Available: https://arxiv.org/abs/2511.10173v2

[19] F. Laitenberger, H. T. Scheuer, H. A. Scheuer, E. Lilienthal, S. You, and R. E. Friedrich, "Cephalometric landmark detection using vision transformers with direct coordinate prediction," *J. Cranio-Maxillofacial Surg.*, vol. 53, no. 9, pp. 1518–1529, Sep. 2025, doi: 10.1016/J.JCMS.2025.05.021.

[20] H. Wu *et al.*, "Cephalometric Landmark Detection across Ages with Prototypical Network," Jun. 2024, Accessed: Jul. 27, 2026. [Online]. Available: https://arxiv.org/pdf/2406.12577

[21] M. A. Khalid *et al.*, "A Benchmark Dataset for Automatic Cephalometric Landmark Detection and CVM Stage Classification," *Sci. Data 2025 121*, vol. 12, no. 1, pp. 1336-, Jul. 2025, doi: 10.1038/s41597-025-05542-3.

[22] C. W. Wang *et al.*, "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.*, vol. 31, pp. 63–76, Jul. 2016, doi: 10.1016/J.MEDIA.2016.02.004.

[23] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Aug. 2019, doi: 10.1109/TPAMI.2020.2983686.

[24] Q. Wu, S. Y. Yeo, Y. Chen, and J. Liu, "Revisiting Cephalometric Landmark Detection from the view of Human Pose Estimation with Lightweight Super-Resolution Head," Sep. 2023, Accessed: Aug. 10, 2026. [Online]. Available: https://arxiv.org/pdf/2309.17143

[25] R. Chen, Y. Ma, N. Chen, D. Lee, and W. Wang, "Cephalometric Landmark Detection by AttentiveFeature Pyramid Fusion and Regression-Voting," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11766 LNCS, pp. 873–881, Aug. 2019, doi: 10.1007/978-3-030-32248-9_97.