Expert Systems With Applications 255 (2024) 124840

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# A two-stage regression framework for automated cephalometric landmark detection incorporating semantically fused anatomical features and multi-head refinement loss

Muhammad Anwaar Khalid [a], Atif Khurshid [a], Kanwal Zulfiqar [b], Ulfat Bashir [b], Muhammad Moazam Fraz [a,*]

[a] National University of Sciences and Technology (NUST), Islamabad, Pakistan
[b] Department of Orthodontics, Riphah International University, Islamabad, Pakistan

ARTICLE INFO

*Keywords:*
Orthodontics
Cephalometric landmark detection
Deep learning
Cascaded CNNs
Feature fusion

ABSTRACT

Accurate identification and precise localization of cephalometric landmarks provide clinicians with essential insights into craniofacial deformities, aiding in the assessment of treatment strategies for improved patient outcomes. The current methodologies heavily depend on the utilization of multiple CNNs for predicting landmark coordinates, which makes them computationally burdensome and unsuitable for translation to clinical applications. To overcome this limitation, we propose a novel, end-to-end trainable, two-stage regression framework for cephalometric landmark detection. In the initial stage, a single neural network is employed to estimate the locations of all landmarks simultaneously, enabling the identification of potential landmark regions. In the second stage, a semantic fusion block leverages the in-network multi-resolution feature hierarchy to produce high-level semantically rich features. These feature maps are then cropped based on coarsely detected landmark locations and concurrent refinement loss is used to fine-tune and refine the landmark locations. The proposed framework demonstrates the potential for enhancing clinical workflow and treatment outcomes in orthodontics. This is achieved through the utilization of a single CNN backbone augmented with multi-resolution semantically fused anatomical features, which effectively enhances representation learning in a computationally efficient manner. The performance of the proposed framework is evaluated on two publicly available anatomical landmark datasets. The experimental results demonstrate that our framework achieves a state-of-the-art detection accuracy of 87.17% within the clinically accepted range of 2 mm. The source code and the pre-trained weights are made publicly available at this https://github.com/manwaarkhd/CEPHMark-Net, promoting reproducibility and enabling further advancements.

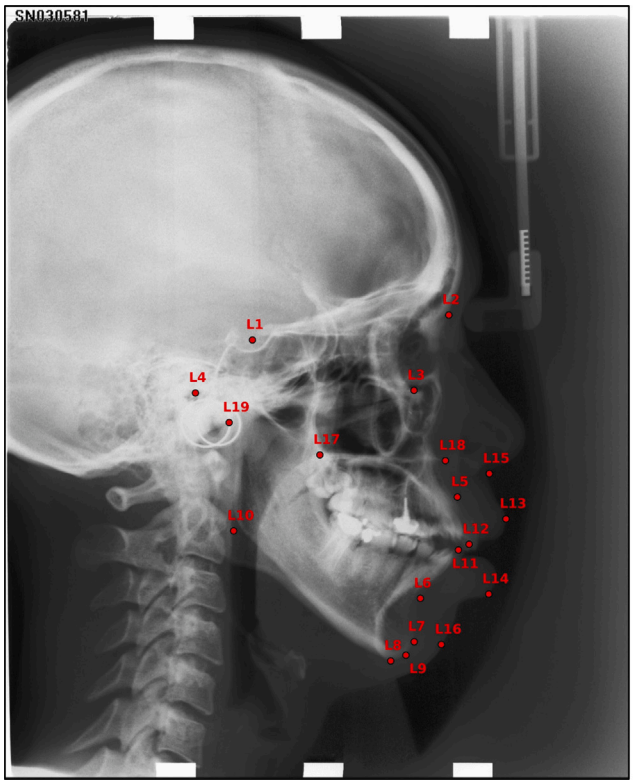
## 1. Introduction

From classic morphometry to contemporary orthodontics, analysis of the intricate relationships among teeth, jaws, maxilla, mandible and the cranial base has been a fundamental examination for orthodontic treatment planning and maxillofacial surgeries (Yue, Yin, Li, Wang, & Xu, 2006). Tracing anatomical landmarks on a two-dimensional (2D) radiograph of the craniofacial region, often referred to as a cephalogram, is a key operation during this analysis (Zeng, Yan, Liu, Zhou, & Qiu, 2021). The quantitative evaluation of these landmarks provides crucial information about the skull and surrounding soft tissue structures (Milošević, Vodanović, Galić, & Subašić, 2022). This information helps orthodontists to accurately classify anatomic facial types, evaluate growth patterns, and obtain a comprehensive picture of patients' craniofacial condition (Wang et al., 2016).

In clinical practice, orthodontists usually identify cephalometric landmarks manually by tracing craniofacial contours. This process, however, is tedious and time-consuming (Kamoen, Dermaut, & Verbeeck, 2001), with even experienced clinicians spending 20-30 min (Qian et al., 2020; Wang et al., 2015) on a single X-ray analysis, creating a significant bottleneck in clinical workflow. Moreover, the accuracy of landmark identification varies with orthodontist's experience, resulting in a risk of inter- and intra-observer variabilities (Singh & Raza, 2022). Since precise estimation of cephalometric landmarks is vital for clinical evaluations and treatment decisions (Lee, Yu, Kim,

* Corresponding author.
*E-mail addresses:* anwar@vision.seecs.edu.pk (M.A. Khalid), atif@vision.seecs.edu.pk (A. Khurshid), kanwal.zulfiqar@riphah.edu.pk (K. Zulfiqar), ulfat.bashir@riphah.edu.pk (U. Bashir), moazam.fraz@seecs.edu.pk (M.M. Fraz).

https://doi.org/10.1016/j.eswa.2024.124840
Received 13 July 2023; Received in revised form 6 June 2024; Accepted 19 July 2024
Available online 22 July 2024




| No. | Landmark | Symbol |
|---|---|---|
| L1 | Sella | S |
| L2 | Nasion | N |
| L3 | Orbitale | Or |
| L4 | Porion | Po |
| L5 | A-point | A |
| L6 | B-point | B |
| L7 | Pogonion | Pog |
| L8 | Menton | Me |
| L9 | Gnathion | Gn |
| L10 | Gonion | Go |
| L11 | Lower incisal incision | LIT |
| L12 | Upper incisal incision | UIT |
| L13 | Upper lip | Ls |
| L14 | Lower lip | Li |
| L15 | Subnasale | Sn |
| L16 | Soft tissue pogonion | `Pog |
| L17 | Posterior nasal spine | PNS |
| L18 | Anterior nasal spine | ANS |
| L19 | Articulare | Ar |

**Fig. 1.** Annotated cephalogram from the IEEE ISBI 2015 challenge dataset (**left**) and the corresponding table containing information about the 19 cephalometric landmarks used in this study (**right**).

Kim, & Choi, 2020), inaccurate manual analysis can result in discrepancies in measuring various craniofacial parameters (Arik, Ibragimov, & Xing, 2017). Therefore, it is imperative to develop an automatic cephalometric landmark detection system that can identify craniofacial landmarks accurately, reliably, and rapidly.

In recent years, significant strides have been made in facial landmark detection through various deep learning-based approaches (Bulat & Tzimiropoulos, 2018; Hou, Wang, Cheng, & Gong, 2015; Sun, Wang, & Tang, 2013; Zhu, Shi, & Gao, 2019). These advancements highlight the effectiveness of advanced, AI-powered learning methods, particularly (Arik et al., 2017; Kwon, Koo, Park, & Cho, 2021; Lee et al., 2020; Zeng et al., 2021), which have demonstrated remarkable results, surpassing traditional machine-learning approaches. However, these approaches utilized a stand-alone CNN model for each landmark without any feature-sharing mechanism. While effective for datasets with a small number of landmarks, this approach becomes increasingly compute and memory-intensive as the number of landmarks increases, thereby limiting their clinical applicability. Additionally, training each CNN model independently results in an approach that is far from end-to-end learning (Oh, Oh, Lee, et al., 2020). Furthermore, landmarks have an inherent graphical structure, where the location of each landmark holds significant importance relative to others. The independent predictions by individual CNNs can introduce uncertainty into the overall estimation process. Given these challenges, there is a need to design a framework that shares a single CNN backbone for detecting multiple landmarks simultaneously, while incorporating an inter-model communication mechanism that allows to learn from each other's predictions and correct errors in real time.

In this research, we approach cephalometric landmark detection as a multi-level regression problem and propose a two-stage detection framework, that follows a coarse-to-fine detection strategy. Unlike other approaches, the proposed framework consists of two modules, integrated within a unified and end-to-end trainable CNN architecture that work in tandem to provide accurate landmark detection. The modules share a single backbone neural network, as a feature extractor, which provides high-dimensional, semantically-rich features to both modules. The first module extracts low-resolution yet semantically strong features from the backbone network and uses them to simultaneously regress coordinates for all landmarks. This approach enables the framework to leverage both global hard/soft tissue characteristics and geometric landmark relations in a unified manner. The task of detecting anatomical landmarks directly within clinical precision range involves a highly non-linear mapping function (Pfister, Charles, & Zisserman, 2015) and becomes even more challenging to learn when the training set is limited in size. To address this issue, prior studies have attempted to refine coarse landmark predictions by using image patches cropped around the directly estimated landmarks. However, relying solely on image patches may not be sufficient, as they only retain information about intensity patterns. Instead, we utilize a cropping mechanism (He, Gkioxari, Dollár, & Girshick, 2018) to extract high-dimensional features from multi-resolution feature tensors generated by the backbone network during the forward pass. To bridge the semantic gap between extracted features, we employ a semantic fusion block that integrates high-resolution, semantically weak features with low-resolution, semantically strong features. This process yields a single, high-level feature map with fine resolution, which is used by the second module to refine the directly estimated landmark coordinates. By exploiting the feature hierarchy of the backbone network, our proposed framework can capture both local and global context information, leading to more accurate and robust landmark detection.

The salient features of our proposed framework are summarized as follows:

- A single, end-to-end trainable, multi-head CNN architecture with a backbone feature extractor. This design allows for the reuse of multi-resolution feature maps from different layers of the backbone network, computed during the forward pass, without additional computational cost.

- Joint learning of both modules that enables the framework to leverage both global hard/soft tissue characteristics and geometric landmark relations in a unified manner.
- A semantic fusion block that effectively leverages feature maps from multiple blocks of the backbone network to generate semantically rich features. Subsequently, the refinement module extracts anatomically relevant features to further refine the initially coarsely predicted landmark locations, with the help of a multi-head refinement loss.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the existing methods in anatomical landmark detection, highlighting their limitations. In Section 3, we present a detailed description of our proposed framework, explaining the key components and their functionality. In Section 4, we report the results of extensive experiments on multiple datasets, including a comparison with state-of-the-art approaches. Finally, we conclude our study in Section 7 by summarizing our main contributions and highlighting future research directions.

## 2. Related work

The following section presents a comprehensive review of existing research on automatic cephalometric landmark detection, where we identify gaps and challenges in the current state-of-the-art.

### 2.1. Traditional image processing approaches

Since the turn of the century, the field of cephalometry has seen a tremendous surge in research, fueled by the ever-growing need for accurate and robust landmark detection. Cardillo and Sid-Ahmed (1994) proposed a landmark recognition algorithm using grey-scale mathematical morphology and template matching. Forsyth and Davis (1996) devised a two-stage approach which first detected all candidate landmark regions using rough and fine feature appearances, and then used spatial image features to choose optimal candidate points. However, the approach was unable to identify the landmarks where a defining structure is not located. Considering this, Grau, Alcaniz, Juan, Monserrat, and Knoll (2001) introduced a novel template-matching approach that incorporates both edge detection and contour segmentation operators, resulting in improved detection performance. Taking inspiration from the dual-phase methodology by Forsyth and Davis (1996), El-Feghi, Sid-Ahmed, and Ahmadi (2004) implemented a neuro-fuzzy system to obtain an initial estimate of landmark locations, followed by a template-matching algorithm to refine their positions within the designated search area. On the other hand, Mohseni and Kasaei (2007) estimated landmark positions using an affine transformation and subsequently refined them using edge-detection, image histogram, and curve-fitting schemes. Although these studies made significant strides in automating cephalometric landmark detection, however, the lack of a benchmark dataset hindered their comparability and reproducibility.

### 2.2. Machine learning approaches

The emergence of machine learning was a beacon of hope in the midst of a challenge to improve the accuracy of cephalometric landmark detection. Catalyzing the shift, the IEEE International Symposium on Biomedical Imaging (ISBI) rose to the occasion by hosting the Automatic Cephalometric Landmark Detection challenge for the Diagnosis in Cephalometric X-ray Images in both 2014 and 2015. The only thing left was a standard dataset, and Wang et al. (2015) completed that missing piece of the puzzle by presenting the first cephalometric dataset, featuring 400 high-resolution X-ray images. During both challenges, the leading solutions were dominated by random forest-based methods where initial landmark positions are estimated using a regression-voting scheme, and later on, the total votes from the region proposals are optimized using a graphical shape model. Among them, Chu, Chen, Nolte, and Zheng (2014) designed a two-phase algorithm with random forest-based landmark detectors and a sparse shape composition model to modify erroneously detected landmarks. Chen and Zheng (2014) proposed a data-driven approach to estimate landmark positions using geometric relationships and gaussian probability maps. Mirzaalian and Hamarneh (2014) used a random forest decision classifier with a pictorial structure algorithm to predict landmarks and added joint distribution of landmark pairs as a regularization term for global optimization. Vandaele, Marée, Jodogne, and Geurts (2014) used an extremely randomized trees-based algorithm for binary pixel classification to determine the landmark position based on the median of the positively classified pixels. Besides, Ibragimov, Likar, Pernuš, and Vrtovec (2014) utilized Haar-like features to detect candidate landmark points, which were then optimized using game-theoretic techniques by combining intensity appearance and shape models. Their method was the only one that achieved a detection rate of over 70% within a 2 mm precision range, earning them first place in the ISBI Grand Challenge 2014. Two years later, Lindner and Cootes (2015) significantly improved detection accuracy by incorporating random forest regression voting within a constrained local model. Although the ISBI grand challenges expanded the horizons of automated cephalometric research, the goal of practical clinical application remained distant, since the most successful team could achieve only 74.94% detection rate within the acceptable precision range.

### 2.3. Deep learning approaches

The integration of artificial intelligence brought about a paradigm shift in the healthcare industry. In response, researchers have turned to the cutting-edge capabilities of deep learning to propel cephalometric research and practice to new heights. Lee, Park, and Kim (2017) pioneered the use of deep learning techniques in addressing the challenge of detecting cephalometric landmarks in dental X-ray images. They approached the problem by considering the x- and y-coordinates of all landmarks as independent variables and subsequently constructed 38 CNN-based regression systems to predict these coordinate variables. On the other hand, Arik et al. (2017) proposed a framework that employed 19 CNNs to output probabilistic estimations of different landmark locations, which were then refined using a shape-based model. Although these approaches demonstrated promising results, their performance was constrained by the lack of advanced CNN architectures. In response, Qian, Cheng, Tao, Lin, and Lin (2019) introduced CephaNet, the first Faster RCNN-based method, which utilizes a multi-task loss to reduce intra-class variations and a two-stage repair strategy to eliminate superfluous or undetected landmarks. Similarly, Lee et al. (2020) proposed a two-stage approach that initially extracts potential regions of interest (ROIs) for every landmark, and subsequently employs a set of Bayesian CNNs to estimate the precise landmark location within the extracted region. On the other hand, Zeng et al. (2021) approached cephalometric landmark detection as a multi-stage regression problem, devising a cascaded three-stage CNN structure with a coarse-to-fine detection strategy. Transitioning to recent advancements, heatmap regression methods have emerged as promising alternatives for landmark detection. For instance, Chen, Ma, Chen, Lee, and Wang (2019) introduced an Attentive Feature Pyramid Fusion module (AFPF) designed specifically for heatmap-based landmark detection, aiming to shape high-resolution and semantically enhanced fusion features by integrating heat maps and offset maps for pixel-wise regression-voting. Similarly, Oh et al. (2020) presented a Fully Convolutional Neural Network (FCN) based framework, incorporating a Local Feature Perturbator (LFP) and an Anatomical Context loss (AC loss) to facilitate learning of anatomical context by leveraging spatial relationships between landmarks. Kwon et al. (2021) developed a multi-stage probabilistic approach for landmark detection, utilizing DeepLabv3 to generate heatmaps representing probability density functions for all

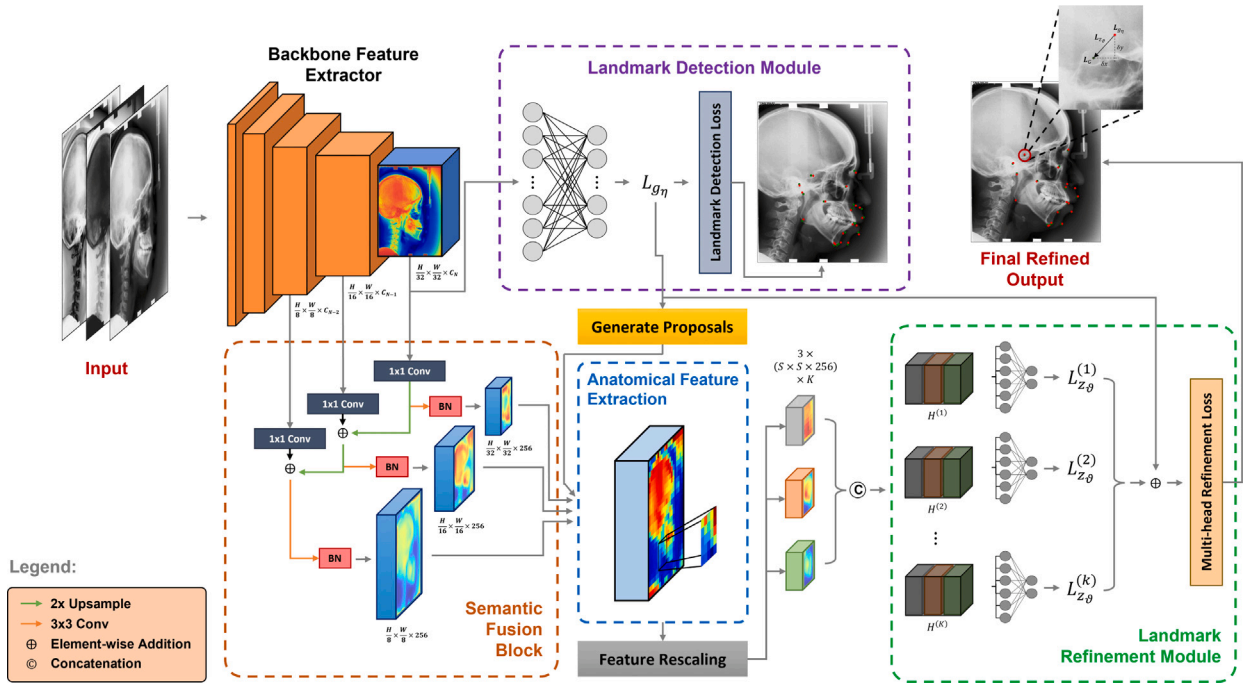



**Fig. 2.** Schematic representation of the proposed framework for landmark detection in cephalometric images. The framework consists of a backbone feature extractor, a landmark detection module (LDM), and a landmark refinement module (LRM), working synergistically to accurately localize anatomical landmarks. The figure includes a patient's cephalogram along with the predictions of the LDM and LRM, clearly illustrating the significant improvements achieved by the refinement module in refining the coarse predictions made by the detection module, leading to highly accurate and refined landmark localizations.

landmarks. This method involves the initial detection of all landmarks using a single network, followed by individual refinement of each landmark using high-resolution cropped images. Notably, each refinement step employs a separate CNN model, to enhance the localization precision by incorporating local details of anatomical structures.

The strides made by deep learning-based approaches in cephalometric landmark detection are commendable. However, current state-of-the-art methods, including both co-ordinate and heatmap-regression based techniques, still face significant limitations hindering their application in clinical settings. While heatmap regression methods demonstrate promising advancements, they also pose challenges. One major limitation is their reliance on a one-to-one mapping between the number of landmarks and the CNN models employed. Although this approach may feasible for datasets with a limited number of landmarks, it escalates computational and memory demands as the landmark count grows. Moreover, the independent training of CNN models lacks necessary inter-model communication, hindering their ability to learn the intricate relationships between landmarks and the surrounding image features. Therefore, a novel framework is required that can address these limitations by utilizing a single, shared CNN backbone for detecting multiple landmarks simultaneously, while incorporating an inter-model communication mechanism that allows models to learn from each other's predictions and correct errors in real time.

## 3. Method

In this section, we present our proposed framework for automated cephalometric landmark detection, which consists of three modules: the Backbone Feature Extractor, the Landmark Detection Module, and the Landmark Refinement Module. We also discuss two different training strategies that allow for sharing convolutional layers across the modules.

### 3.1. Problem formulation

In this study, our objective is to develop an automated framework that accurately predicts the positions of anatomical landmarks, based solely on dental X-ray images. In this context, we formally define the cephalometric landmark detection problem as follows: Let $\mathcal{X} \in \mathbb{N}^{H\times W\times 3}$ represent the set of X-ray images, where $W$ and $H$ denote their width and height, respectively. Furthermore, let $\mathcal{Y} \in \mathbb{R}^{K\times 2}$ represent the coordinate space for $K$ number of anatomical landmarks. Therefore, given an input image $X \in \mathcal{X}$, the task of predicting the corresponding landmarks $Y \in \mathcal{Y}$ can be viewed as learning a nonlinear mapping function $\Phi$, which we can express as follows:

$$\Phi : X \rightarrow Y \tag{1}$$

This mapping function $\Phi$ is highly non-linear (Pfister et al., 2015; Tompson, Jain, LeCun, & Bregler, 2014) and solving it directly becomes even more challenging when the training data is limited in size. Consequently, previous studies such as (Arik et al., 2017; Chu et al., 2014; El-Feghi et al., 2004; Forsyth & Davis, 1996; Lee et al., 2020) have adopted a multi-stage detection approach to address the landmark regression problem. This approach also inspired the design of our proposed framework, which utilizes a two-stage regression model for cephalometric landmark detection, following a coarse-to-fine prediction strategy.

Our proposed framework is a multi-head CNN architecture that shares a single backbone neural network, as illustrated in Fig. 2. This shared backbone acts as a feature extractor, providing high-dimensional and semantically rich features to other modules. Moreover, it facilitates inter-module communication, allowing modules to learn from each other's predictions and align themselves accordingly. This unique feature of our proposed framework sets it apart from previous approaches and makes it not only end-to-end trainable but also more accurate and robust.

### 3.2. Framework architecture

In the following subsections, we provide a detailed description of each module and demonstrate how they synergize to enable robust and accurate landmark detection.

### 3.2.1. Backbone feature extractor

The proposed framework employs a convolutional neural network $f_\theta(\cdot)$, parameterized by weights $\theta$, as the feature extractor to transform input X-ray images into high-dimensional feature representations. Our framework offers the flexibility to use various network architectures as a backbone without any constraints. Specifically, we extract feature activation outputs from the last convolution layer of each block in the backbone network to obtain our reference set of feature maps, with the deepest layer being chosen due to its possession of semantically strong features. These feature maps are denoted by $F_b \in \mathbb{R}^{H_b \times W_b \times C_b}$, where $H_b$, $W_b$, and $C_b$ represent the height, width, and channels of feature map $F_b$, respectively, while $b$ represents the block number.

The initial blocks of the backbone network produce feature maps with large spatial dimensions, which can be memory-intensive and computationally expensive to process in later stages of the network. Additionally, these feature maps often do not contain significant semantic information. Therefore, to optimize computational complexity and avoid memory constraints, we limit ourselves to using the feature maps from the last three blocks of the backbone network. These feature maps have smaller spatial dimensions and more meaningful semantics, providing a better representation for subsequent modules. We define the backbone feature extractor as:

$$f_\theta : X \rightarrow F_{n-2}, F_{n-1}, F_n \tag{2}$$

where $n$ denotes the total number of blocks in the backbone network.

### 3.2.2. Landmark detection module

The landmark detection module is a crucial component of our framework, designed to detect all anatomical landmarks in an input X-ray image simultaneously. By considering all landmarks together, this module effectively captures global hard/soft tissue characteristics and exploits the existing geometric relationships between landmarks (Kwon et al., 2021). To achieve this, we use the feature map $F_n$ from the deepest layer of the backbone network as input to a detection network $g_\eta(\cdot)$, parameterized by weights $\eta$, which maps the feature representations to coarse landmark locations. The feature map $F_n$ is flattened to a feature vector $\mathbf{F}'_n \in \mathbb{R}^{H_n \times W_n \times C_n}$ and passed through a fully connected layer with $2K$ neurons, where $K$ represents the number of landmarks. The output of $g_\eta(\cdot)$ is a vector $\mathbf{L}g_\eta \in \mathbb{R}^{2K}$, which represents the predicted x- and y-coordinates of all the landmarks, as follows:

$$\mathbf{L}_{g_\eta} = g_\eta(F_n) = [\hat{x_1}, \hat{y_1}, \ldots, \hat{x_K}, \hat{y_K}] \tag{3}$$

Here, $\hat{x_i}$ and $\hat{y_i}$ represent the predicted x- and y-coordinates of the $i$th landmark, respectively. The weights $\eta$ of the detection network are optimized using mean squared error (MSE) as the loss function, which is defined as:

$$\mathcal{L}_{g_\eta} = \frac{1}{K} \sum_{i=1}^{K} \left\| \mathbf{L}_G^{(i)} - \mathbf{L}_{g_\eta}^{(i)} \right\|^2 \tag{4}$$

Here, $\mathbf{L}_G^{(i)} = [x_i, y_i]$ and $\mathbf{L}g_\eta^{(i)} = [\hat{x_i}, \hat{y_i}]$ represent the ground truth and predicted locations of $i$th landmark, respectively. The objective is to find the optimal values of $\eta$ that minimize the distance between the ground truth and predicted landmarks.

### 3.2.3. Semantic fusion block

The backbone network generates feature maps at multiple resolutions, each with different levels of semantic information. To bridge the semantic gap between these feature maps, we introduce a semantic fusion block (SFB) that leverages feature maps from multiple blocks and fuses them to produce semantically rich features. This approach ensures rich semantic information at all levels and can be efficiently implemented with a single input image scale.

Specifically, we attach a $1 \times 1$ convolution layer to the lowest resolution feature map $F_n$ to produce the coarsest features, followed by an up-sampling layer which increases its spatial dimensions by a factor of 2. The up-sampled feature map is then merged with $F_{n-1}$, by element-wise addition, which has undergone a $1 \times 1$ convolution to reduce channel dimensions. This process is repeated until the finest resolution map is obtained.

To mitigate the aliasing effect caused by up-sampling, we apply a $3 \times 3$ convolution on each merged map to generate the final feature maps. The output feature maps from the SFB are denoted by $P_{n-2}$, $P_{n-1}$, and $P_n$, corresponding to the feature maps obtained from the last three blocks of the backbone network, respectively. By combining information from multiple blocks, the SFB can capture both fine-grained, local feature semantics from high-resolution feature maps and coarse, global feature semantics from low-resolution feature maps. This helps refine the landmark localization and produce more accurate results.

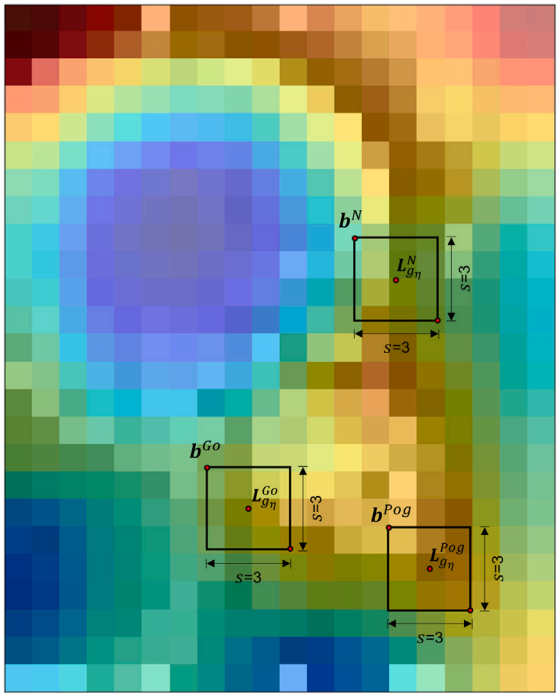



**Fig. 3.** Illustration of candidate region proposals of size $s = 3$ for Gonion, Pogonion and Nasion. The heatmap image shows the feature map from the last block of the backbone network. For each landmark, the region proposal $\mathbf{b}^k$ is generated by considering $\mathbf{L}^k_{g_\eta}$ as the center with a size $s$ in both directions.

### 3.2.4. Landmark refinement module

The landmark refinement module utilizes a region proposal approach to further refine the landmark locations predicted by the detection network. To achieve this, we first generate candidate region proposals using the predicted landmark locations as follows:

$$\begin{aligned} \mathbf{b}_{x_1} &= \mathbf{L}^{(x)}_{g_\eta} - (s/2)\ ;\ \mathbf{b}_{y_1} = \mathbf{L}^{(y)}_{g_\eta} - (s/2) \\ \mathbf{b}_{x_2} &= \mathbf{L}^{(x)}_{g_\eta} + (s/2)\ ;\ \mathbf{b}_{y_2} = \mathbf{L}^{(y)}_{g_\eta} + (s/2) \end{aligned} \tag{5}$$

Here, $\mathbf{L}g_\eta^{(x)}$ and $\mathbf{L}g_\eta^{(y)}$ represent the predicted landmark coordinates along $x$ and $y$ axes, respectively, while $s$ denotes the size of the region proposal.

For each cephalogram, a total of $K$ region proposals are generated for a given value of $s$, as demonstrated in Fig. 3. Since the feature maps produced by SFB have varying resolutions, different values of $s$ are selected for each $P_i$. This strategy ensures that the proposals cover relatively different contextual regions across the feature maps. The proposal $\mathbf{b}^k_i$ of size $s_i$ is then used to crop the feature map $P_i$, extracting the region of interest for $k$th landmark. The cropped feature map is denoted by $D^k_i \in \mathbb{R}^{s_i \times s_i \times C_i}$, where $C_i$ represents the number of channels in $P_i$. This process is repeated for all the feature maps generated by

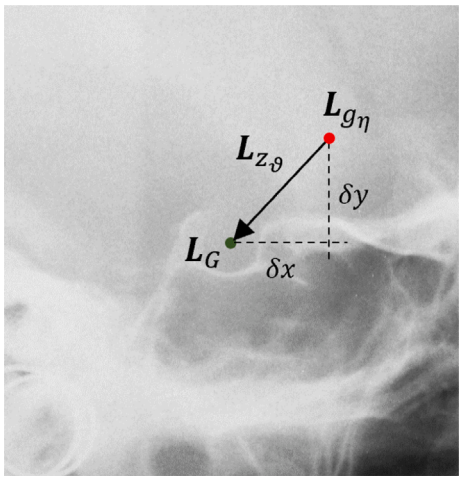



**Fig. 4.** Illustration of the landmark refinement process for the landmark Sella. The objective of the refinement head is to predict refinement values $\delta x$ and $\delta y$ for each landmark such that the predicted landmark $\mathbf{L}_{g_\eta}$ approaches the ground truth landmark $\mathbf{L}_G$ as closely as possible.

SFB. Subsequently, the cropped feature maps are resized to a single spatial resolution $S$, allowing them to be concatenated to form a single high-level feature map $H^{(k)} \in \mathbb{R}^{S\times S\times C_{n-2}+C_{n-1}+C_n}$ for the $k$th landmark. The concatenation process combines feature maps from all levels of the SFB, enabling the model to capture multi-scale contextual information.

The landmark refinement module uses a refinement head $z_\vartheta^k(\cdot)$, parameterized by weights $\vartheta$, to predict the refinements $\delta x^k$ and $\delta y^k$ for the $k$th landmark based on the high-level feature map $H^{(k)}$. This can be represented as follows:

$$z_\vartheta^k(H^{(k)}) = [\delta x^k, \delta y^k] = \mathbf{L}_{z_\vartheta}^k \in \mathbb{R}^2 \tag{6}$$

Here, $\mathbf{L}_{z_\vartheta}^k$ represents the predicted refinements against $k$th landmark. Fig. 4 demonstrates the refinement process for the landmark Sella, where $z_\vartheta^S$ predicts refinement values, bringing the $\mathbf{L}_{g_\eta}^S$ closer to the $\mathbf{L}_G^S$. The weights of the refinement head are optimized by minimizing the mean squared error between the ground truth landmark $\mathbf{L}_G^K$ and the sum of predicted landmark $\mathbf{L}_{g_\eta}^k$ and its corresponding refinements $\mathbf{L}_{z_\vartheta}^k$. This is expressed as the following loss function for the $k$th refinement head:

$$\mathcal{L}_{z_\vartheta}^k = \left\| \mathbf{L}_G^k - (\mathbf{L}_{z_\vartheta}^k + \mathbf{L}_{g_\eta}^k) \right\|^2 \tag{7}$$

The combined loss of all the heads is referred to as the multi-head refinement loss, which integrates the refinement performance of each head into a unified optimization objective. This enables the network to collectively refine the predicted landmarks and enhance the overall localization accuracy.

### 3.3. Training strategy

The proposed framework comprises a detection network and refinement heads, all connected to a shared backbone. If trained separately, these modules will modify the shared convolutional layers differently, making it challenging to learn an optimal feature representation. To avoid this, we need to design a training strategy that allows for sharing convolutional layers between the modules, rather than learning two separate networks. We experimented with the following strategies for training the modules with a shared backbone.

#### 3.3.1. Alternate training

In this approach, we first train the landmark detection module, attached to the backbone feature extractor, and use the predicted locations to generate landmark region proposals. The computed region proposals are stored for subsequent use as inputs to the refinement module. We then freeze the weights of the backbone network and train the refinement heads using pre-computed proposals to improve the accuracy of the predicted locations. This approach allows us to train the two modules separately while still sharing convolutional layers, reducing the overall computational cost of training the framework. However, it also results in a slower convergence rate due to the iterative nature of the approach.

#### 3.3.2. Joint training

In this approach, the detection network and refinement heads are merged into a single network, as depicted in Fig. 2. In each iteration, candidate region proposals are generated during the forward pass and treated as fixed, pre-computed proposals for training the refinement heads. The backward propagation is performed as usual, with signals propagated from both the detection loss and the refinement loss being combined for the shared layers (Ren, He, Girshick, & Sun, 2015). Specifically, the overall loss function $\mathcal{L}$ is a weighted sum of the detection loss and the refinement loss for all $K$ landmarks, given by:

$$\mathcal{L} = \alpha \mathcal{L}_{g_\eta} + \beta \sum_{k=1}^{K} \mathcal{L}_{z_\vartheta}^k \tag{8}$$

Here, $\mathcal{L}_{g_\eta}$ is the loss function for the detection module, and $\mathcal{L}_{z_\vartheta}^k$ is the loss function for the refinement module for the $k$th landmark. The hyperparameters $\alpha$ and $\beta$ weight the contributions of the detection and refinement losses, respectively. By merging the two modules, the joint training approach allows for the refinement module to improve the detection module's feature representation, leading to better accuracy in both detection and refinement tasks. The shared backbone in this approach allows for more efficient use of computational resources, reducing the overall training time required to achieve good performance.

### 3.4. Implementation details

In the following subsections, we delve into the implementation details of our framework, including the pre-processing steps and the selection of hyperparameters for training.

#### 3.4.1. Experiment environment

Our proposed framework is implemented in TensorFlow,[1] a comprehensive open-source platform for machine learning, and trained on a system equipped with NVIDIA GeForce RTX 3060 Lite GPU, with 12 GB of RAM, running Ubuntu 22.04.1 operating system.

#### 3.4.2. Preprocessing

Since the training dataset only contains a limited number of images, *i.e.*, 150, it is necessary to augment the data to improve the model's generalization ability. Therefore, we apply various augmentation techniques, as described in Section 4.2. Additionally, we normalize the pixel values in each sample to ensure that the data has zero mean and unit variance, which is a common practice in deep learning to stabilize the training process and improve the convergence speed. To achieve this, we convert the pixel values from the range $\{P \in \mathbb{N} \mid 0 \le p \le 255\}$ to the range $\{P' \in \mathbb{R} \mid 0 \le P' \le 1\}$ using the following formula where $s$ and $\gamma$ are set to be 255 and 0, respectively:

$$P' = \frac{P}{s} + \gamma \tag{9}$$

Moreover, due to the high original spatial dimensions of the cephalograms, it is not feasible to use them in their original form. Therefore, we downscale the images by a factor of 3 using bi-linear interpolation, which preserves the image's overall shape and reduces the aliasing artifacts. The new dimensions of images after re-scaling are

[1] https://www.tensorflow.org/

$800 \times 620 \times 3$. The corresponding landmarks are transformed using the following formula to adjust to the new dimensions:

$$\mathbf{L}'_x = \frac{\mathbf{L}_x}{\epsilon} \; ; \; \mathbf{L}'_y = \frac{\mathbf{L}_y}{\epsilon} \tag{10}$$

Here, $\mathbf{L}x$ and $\mathbf{L}y$ are the original $x$ and $y$ coordinates of the landmarks and $\epsilon$ is the scaling factor, which is set to 3 in this case.

#### 3.4.3. Hyper-parameters

The network weights are independently initialized using the Xavier policy (He, Zhang, Ren, & Sun, 2015b) and updated through stochastic gradient descent (SGD) (Ruder, 2017) with an initial learning rate of 0.01 over a total of 500 epochs. To balance the training efficiency and model performance, we employ a batch size of 4 during the training process.

## 4. Experiments

In this section, we present the experimental settings of our framework, including the dataset utilized for development and training, the evaluation metrics employed to assess performance and implementation details.

### 4.1. Dataset

The choice of dataset plays a vital role in the development and evaluation of AI algorithms. For this study, we utilized the publicly available cephalometric dataset[2] by Wang et al. (2015), consisting of 400 high-resolution X-ray images collected from patients ranging in age from 6 and 60 years. Each image has spatial dimensions of $1935 \times 2400$ pixels, with a spatial resolution of 0.1 mm/pixel in both directions. An annotated cephalogram, featuring all 19 anatomical landmarks, is presented in Fig. 1.

Two experienced orthodontists meticulously annotated each cephalogram by identifying 19 anatomical landmarks. To account for inter-observer variability, both orthodontists reviewed annotations, and the average of both reviews was considered the ground truth. To ensure the objectivity of our experimental results and to facilitate a comparison with prior studies, we employed the same 150 images for training as used in the ISBI Grand Challenge 2015. The remaining 250 images are reserved for evaluation, and further partitioned into two distinct subsets, Test1 and Test2, with the former serving as our validation set for assessing the accuracy of our method during the development phase and the latter as our test set for the final evaluation of our proposed method.

### 4.2. Data augmentation

Unlike other areas of computer vision, the datasets for medical image analysis are often limited in size, as the process of obtaining ground-truth labels from clinical experts is resource-intensive and time-consuming. Similarly, publicly available cephalometric datasets are also scarce and have a limited number of images (*e.g.* cephalometric landmark detection dataset by Wang et al. (2015) has only 150 radiographs for training). Consequently, training deep learning models on such a small amount of data can lead to overfitting and result in poor performance in clinical applications. To mitigate this issue, data augmentation has emerged as a powerful method (Shorten & Khoshgoftaar, 2019). Previous studies, such as Oh et al. (2020), Qian et al. (2020), Zeng et al. (2021), have also employed data augmentation techniques, however, the transformations used in these studies may not be ideal for radiographic images. Given that cephalometric radiographs hold rich and complex morphometric information, it is important to use specialized data augmentation techniques, as demonstrated by Maini and Aggarwal (2010), to ensure that the synthetic data remains clinically relevant and useful for training robust models. In our study, we employ a diverse range of specialized image transformations, as elaborated below.

#### 4.2.1. Photometric transformations

- **Adaptive histogram equalization** We utilized the contrast-limited adaptive histogram equalization (CLAHE) technique, proposed by Koonsanit, Thongvigitmanee, Pongnapang, and Thajchayapong (2017), with a grid size of 128. This technique has proven effective in amplifying local contrast while preserving fine details and textures in the radiographs.
- **High-pass filtering** We accentuate the sharp changes in intensity levels of a cephalogram using high-frequency emphasis filtering (HEF) with a radius $r$ randomly drawn from $[0, 1]$.
- **Unsharp masking** To enhance contrast and sharpness as done by Deng (2010), we utilized a linear filter that selectively amplifies the high-frequency content of radiographs. The applied method can be represented by Eq. (11).

$$I_{sharp} = I_{orig} + \alpha * (I_{orig} - I_{orig} * F) \tag{11}$$

  where $F$ is a linear filter and $\alpha$ is the sharpness amount.
- **Solarization** We utilized the solarization method to mitigate tone overexposure in radiographs and enhance the visibility of complex craniofacial structures.
- **Inversion** To further increase the diversity of our training data, we employed the inversion technique, which reverses the intensity values of radiographs making light areas dark and vice versa.

#### 4.2.2. Geometric transformations

- **Translation** We translated a radiograph by a translation factor $f_t$ in both horizontal and vertical direction with $\delta$ as a soft margin to prevent landmarks from being translated to the edge of the image.
- **Cropping** In order to ensure that all anatomical landmarks are present in the cropped region, a craniofacial bounding box with a margin $\delta$, randomly selected from $[128, 192]$, was used to crop the radiograph. The corresponding landmarks are then translated using:

$$\mathbf{L}^{(i)}_{(x,y)} = \mathbf{L}^{(i)}_{(x,y)} - \text{bbox}^{(i)}_{(x,y)} \tag{12}$$

  where $\mathbf{L}^{(i)}_{(x,y)}$ are landmark coordinates of image $i$ along x- and $y$-axis, respectively, and $\text{bbox}^{(i)}_{(x,y)}$ represents the top-left corner of the craniofacial bounding box.
- **Reflection** Radiographs are subjected to a random flip, and subsequently, the corresponding landmarks are translated as:

$$\mathbf{L}^{(i)} = \begin{pmatrix} W \\ H \end{pmatrix} - \mathbf{L}^{(i)} \tag{13}$$

  where $\mathbf{L}^{(i)}$ is a vector containing the x- and y-coordinates of the $i$th landmark, and the subtraction is performed element-wise.

#### 4.2.3. Cephalometric transformations

Using the clinical precision range of 2.0 mm, we induced random variations in the ground truth values of landmarks in all directions. The transformation can be represented mathematically as follows:

$$\mathbf{L}^{(i)} = \mathbf{L}^{(i)} - \lambda \tag{14}$$

where $\lambda \rightarrow \mathbb{N}^{19\times 2}$ and $\lambda_{(i)} \in [-\delta, \delta]$. The value of $\delta$ is set to 10, serving as the optimal threshold for our particular experiment.

We first performed data augmentation by randomly applying the aforementioned techniques on-the-fly during training. However, this led to severe fluctuations in the learning curve, adversely affecting the

[2] https://figshare.com/s/37ec464af8e81ae6ebbf

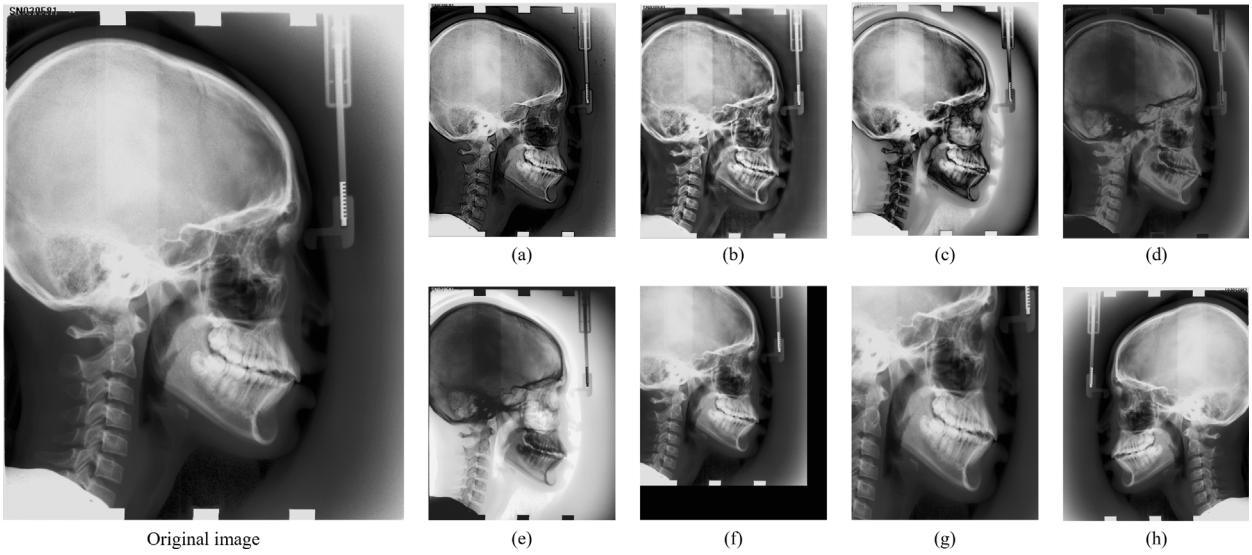



**Fig. 5.** A visual representation of the transformative power of our image augmentation techniques on cephalograms. The left-most column shows the original image, while the remaining columns demonstrate the effects of (a) Adaptive histogram equalization, (b) High-pass filtering, (c) Unsharp masking, (d) Solarization, (e) Inversion, (f) Random translation, (g) Random cropping, and (h) Random reflection.

training process. To overcome this issue, we pre-generated a training set of 4200 augmented images by applying these techniques to the original 150 training radiographs. A comprehensive overview of the visual effects of all the data augmentation techniques applied in our study can be found in Fig. 5.

### 4.3. Evaluation metrics

The effectiveness of our proposed method is evaluated based on two main statistical measures.

- **Mean radial error (MRE)** The first evaluation criterion is the mean radial error, which measures the average distance between the predicted and ground truth landmarks. The radial error $R_i$ for the $i$th landmark is calculated as the euclidean distance between the predicted and ground truth positions. The Mean Radial Error (MRE) is calculated as follows:

$$MRE = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{15}$$

  where $N$ is the total number of landmarks and $R_i$ is the radial error for the $i$th landmark.
- **Success detection rate (SDR)** The second evaluation criterion is the Success Detection Rate (SDR) with a precision range of 2.0 mm, 2.5 mm, 3.0 mm, and 4.0 mm. For each landmark, a reference location is marked by orthodontists as a single pixel. A detected landmark is considered successful if the absolute difference between the detected and ground truth landmark positions is within the clinically accepted range, and otherwise, it is considered a detection failure. The success detection rate $p_z$ with precision less than $z$ is calculated as follows:

$$p_z = \frac{\{\#i : \|L_d(i) - L_g(i)\| < z\}}{\#\Omega} \times 100 \tag{16}$$

  where $L_d$ and $L_g$ represent the positions of the detected and ground truth landmarks, respectively, $z$ denotes the precision range, and $\#\Omega$ is the total number of detections made.

## 5. Results

We evaluated the performance of our proposed cephalometric landmark detection framework on two test datasets, Test-1 and Test-2, comprising 150 and 100 cephalometric images, respectively. Table 1 summarizes the results, including the mean radial error (MRE) with standard deviation (SD) for all 19 landmarks on both test sets, as well as the success detection rate (SDR) within four clinical ranges (2.0 mm, 2.5 mm, 3.0 mm, and 4.0 mm) for each landmark. The average MRE with SD across all landmarks for 250 test images is evaluated to $1.51 \pm 0.91$ mm, falling within the clinically accepted precision range of 2.0 mm. Moreover, as evidenced by the smaller MRE and SD values, the results indicate our framework's capability of locating cephalometric landmarks accurately and consistently.

In Test-1, MRE varies from 0.87 mm (L8) to 1.97 mm (L4), with SDR ranging from 63.33% to 95.33% within 2.0 mm neighborhood. On average, all landmarks exhibit an average MRE and SD of $1.23 \pm 0.89$ mm, with SDRs of 87.17%, 90.77%, 95.18% and 98.12% for neighborhoods of 2.0 mm, 2.5 mm, 3.0 mm and 4.0 mm, respectively. Similarly, for Test-2, MRE varies from 0.83 (L15) to 4.17 (L16), with SDR ranging from 29.0% to 96.0% within 2.0 mm neighborhood. The average MRE and SD for all landmarks is $1.78 \pm 0.92$ mm, with SDRs of 75.79%, 84.0%, 89.0% and 94.53% for neighborhoods of 2.0 mm, 2.5 mm, 3.0 mm and 4.0 mm, respectively. Additionally, the framework exhibits robust performance in detecting most landmarks, with SDR values above 90% across both test datasets. Notably, landmarks L1, L8, L9, L11, L12, L15, L17 and L18 consistently yield high SDR values, further attesting to the reliability of our approach.

Furthermore, our framework exhibits impressive accuracy in detecting five different types of landmarks, including Porion, Gonion, Upper Lip, Subnasale, and Articulare, achieving SDRs of 63.0%, 73.50%, 62.67%, 93%, and 80.67%, respectively, within 2.0 mm range. Notably, detecting Gonion presents a significant challenge as reflected by other methods such as (Ibragimov et al., 2014; Oh et al., 2020; Wang et al., 2016), but our framework achieves a breakthrough with an average SDR of 73.50%, 81.17%, 88.84%, and 95.50% across the test sets. Fig. 6 provides a qualitative assessment of our proposed framework, showcasing examples of cephalograms along with their ground truth landmarks, predicted landmarks from the Landmark Detection Module $g_\eta$, and the refinements made by the Landmark Refinement Module $z_\vartheta$.

Overall, the proposed framework performs better on Test-1 compared to Test-2, as evidenced by the lower MRE values and higher SDR values for Test-1 across all neighborhood sizes. However, there are significant deviations in error rates for certain landmarks, such as L6 (B-point), L13 (Upper-lip), and L16 (Soft tissue pogonion), between

**Table 1**
Performance evaluation of the proposed framework on IEEE ISBI 2015 Challenge Test datasets, *i.e.* Test-1 and Test-2. The results are presented in terms of mean radial error (MRE) and successful detection rate (SDR) within 2.0, 2.5, 3.0, and 4.0 mm neighborhoods for all 19 landmarks. MRE is reported in millimetres (mm), and SDR values are reported as a percentage (%).

| | ISBI 2015 Challenge Test 1 dataset | | | | | ISBI 2015 Challenge Test 2 dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MRE ± SD (mm) | SDR (%) | | | | MRE ± SD (mm) | SDR (%) | | | |
| | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm |
| L1 | 1.11 ± 1.03 | 93.33 | 95.33 | 98.00 | 98.67 | 0.93 ± 0.87 | 95.00 | 98.00 | 98.00 | 100.00 |
| L2 | 1.32 ± 1.06 | 85.33 | 90.67 | 92.67 | 95.67 | 1.41 ± 0.89 | 78.00 | 87.00 | 95.00 | 98.00 |
| L3 | 1.25 ± 0.95 | 88.33 | 93.67 | 97.33 | 98.33 | 2.14 ± 0.91 | 75.00 | 81.00 | 87.00 | 90.00 |
| L4 | 1.97 ± 1.29 | 63.33 | 71.33 | 89.67 | 94.67 | 2.41 ± 0.87 | 63.00 | 78.00 | 89.00 | 98.00 |
| L5 | 1.73 ± 0.67 | 67.00 | 76.67 | 94.00 | 97.33 | 1.59 ± 0.82 | 78.00 | 89.00 | 95.00 | 99.00 |
| L6 | 1.18 ± 0.81 | 89.33 | 91.67 | 98.00 | 100.00 | 2.99 ± 1.57 | 32.00 | 48.00 | 59.00 | 84.00 |
| L7 | 1.15 ± 0.76 | 90.67 | 93.33 | 96.67 | 100.00 | 1.13 ± 0.96 | 93.00 | 95.00 | 98.00 | 98.00 |
| L8 | 0.87 ± 0.83 | 93.00 | 95.67 | 97.00 | 98.67 | 0.99 ± 0.86 | 96.00 | 98.00 | 99.00 | 100.00 |
| L9 | 0.91 ± 0.76 | 90.67 | 97.67 | 98.00 | 99.33 | 0.87 ± 0.77 | 95.00 | 98.00 | 98.00 | 98.00 |
| L10 | 1.67 ± 1.07 | 68.00 | 74.33 | 84.67 | 95.00 | 2.27 ± 1.17 | 79.00 | 88.00 | 93.00 | 98.00 |
| L11 | 0.96 ± 0.87 | 95.33 | 95.67 | 98.67 | 98.67 | 1.69 ± 0.73 | 93.00 | 95.00 | 98.00 | 99.00 |
| L12 | 0.91 ± 0.61 | 96.67 | 97.33 | 98.33 | 100.00 | 1.86 ± 0.98 | 91.00 | 96.00 | 98.00 | 98.00 |
| L13 | 1.08 ± 1.04 | 95.33 | 96.00 | 96.00 | 98.67 | 2.73 ± 1.07 | 30.00 | 57.00 | 69.00 | 95.00 |
| L14 | 1.32 ± 0.92 | 95.00 | 95.67 | 98.00 | 99.33 | 2.39 ± 0.98 | 47.00 | 73.00 | 87.00 | 96.00 |
| L15 | 0.93 ± 0.87 | 96.00 | 96.67 | 96.67 | 98.67 | 0.83 ± 0.87 | 90.0 | 96.00 | 98.00 | 100.00 |
| L16 | 1.41 ± 0.77 | 91.33 | 92.00 | 94.67 | 98.00 | 4.17 ± 1.17 | 29.00 | 31.00 | 37.00 | 48.00 |
| L17 | 0.95 ± 0.73 | 95.00 | 97.33 | 98.00 | 99.67 | 0.85 ± 0.53 | 95.00 | 98.00 | 99.00 | 100.00 |
| L18 | 1.19 ± 0.84 | 91.33 | 95.33 | 96.33 | 98.67 | 1.26 ± 0.74 | 91.00 | 95.00 | 96.00 | 98.00 |
| L19 | 1.48 ± 1.11 | 71.33 | 78.33 | 85.67 | 95.00 | 1.38 ± 0.71 | 90.00 | 95.00 | 98.00 | 99.00 |
| **Average** | **1.23 ± 0.89** | **87.17** | **90.77** | **95.18** | **98.12** | **1.78 ± 0.92** | **75.79** | **84.00** | **89.00** | **94.53** |

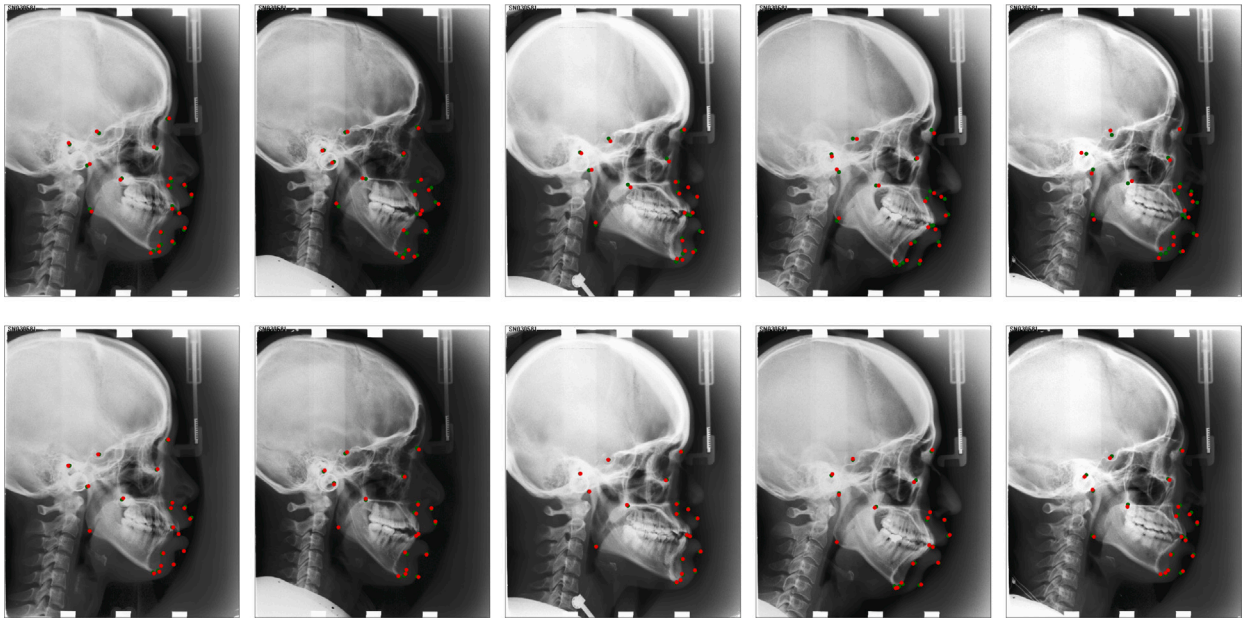


**Fig. 6.** An illustration of the outcomes predicted by our proposed framework. (**top**) presents examples of cephalograms, along with their ground truth landmarks and the predicted landmarks from the Landmark Detection Module $g_\eta$. (**bottom**) exhibits the refinements predicted by the Landmark Refinement Module $z_\vartheta$ for the same set of coarsely predicted landmarks. The cephalograms are arranged based on the performance of $g_\eta$, from the best (**left**) to the worst (**right**). It can be observed that $z_\vartheta$ is efficiently refining the landmarks where the predictions of the $g_\eta$ are the worst (as seen in the right-most cephalogram). These visual results provide compelling evidence for the robustness and accuracy of our proposed framework.

Test-1 and Test-2. In particular, for L16, the MRE value increases up to $4.14 \pm 1.17$ mm in Test-2, resulting in only a 29% SDR for this landmark within the 2.0 mm range. Even within the 4.0 mm range, the SDR is only 48%. The deviations stem from the variability in the annotations of landmarks in the dataset. In fact, the inter-observer variability between the junior and senior experts on the test sets reaches a value of $5.0 \pm 2.80$ mm for L16, indicating significant fluctuation in tracing position by one or both experts. This uncertainty in ground truth landmarks also contributes to the landmark detection module's production of notably inaccurate predictions compared to ground truth. Consequently, the feature-map region cropped from these predictions often lacks either landmark-oriented contextual information or is insufficiently detailed, hindering effective landmark refinement and overall framework performance.

### 5.1. Comparison with state-of-the-art

Our proposed framework demonstrates state-of-the-art performance when evaluated using mean radial error (MRE) and successful detection rate (SDR) within the clinical ranges of 2.0 mm, 3.0 mm, and 4.0 mm on Test-1, while achieving comparable results on Test-2. As shown in Table 2, our framework achieves the highest SDR of

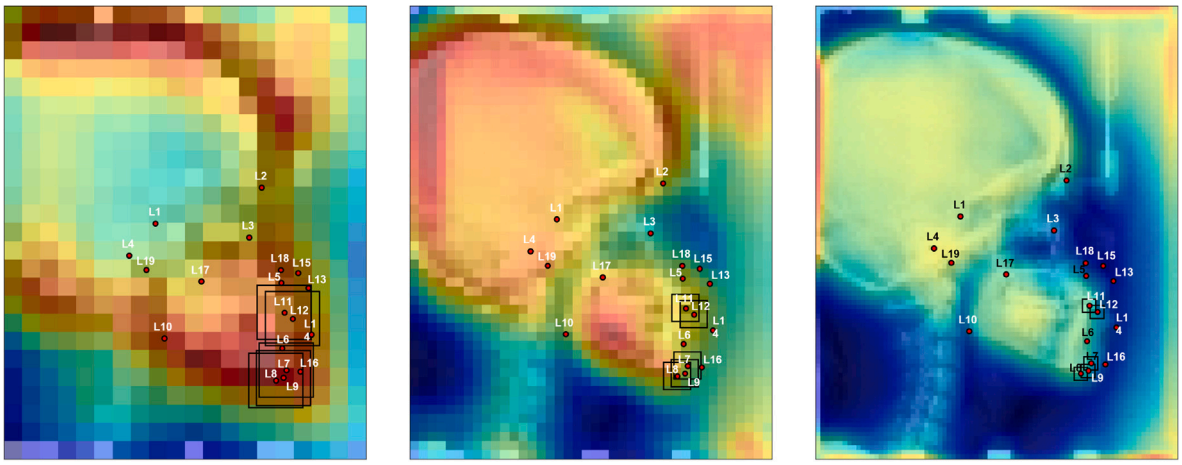



**Fig. 7.** Illustration of the challenge posed by overlapping region proposals for critical landmarks due to the substantial reduction in spatial dimensions in the feature map of the last block. (a) Region proposals surrounding landmarks Menton, Gnathion, and Pogonion exhibit significant overlap, indicated by the high IoU (Intersection over Union) value exceeding 0.8. This overlap presents a hindrance to the refinement module's ability to accurately predict refinements. As a result, the corresponding feature regions cropped for these landmarks are more or less the same, making it difficult for the refinement module to accurately refine them within the same context. (b) and (c) demonstrate that increasing the spatial resolution significantly reduces the overlap among candidate landmark regions.

87.17% among all methods. Furthermore, our framework excels in the 3.0 mm and 4.0 mm clinical ranges with SDR values of 95.18% and 98.12% respectively. These outcomes validate the effectiveness of our framework, which benefits from an efficient design that enables precise and accurate detection and localization of the target landmarks.

On the other hand, it can be observed from Table 2 that our framework has a higher MRE value for the clinical range of 2.0 mm and a slightly smaller SDR value for the clinical range of 2.5 mm compared to Kwon et al. (2021). However, considering the overall performance and the achieved SDR in other clinical ranges, our framework offers promising results for the detection and localization of cephalometric landmarks.

In conclusion, our proposed framework demonstrates the potential to be employed in real-world applications. The achieved high SDR in the critical clinical ranges of 2.0 mm, 3.0 mm, and 4.0 mm signifies its effectiveness, while the slightly higher MRE for the 2.0 mm range can be mitigated by its favorable SDR and comparable results in other ranges.

## 6. Discussion

The accurate localization of anatomical landmarks in cephalometric images is essential for orthodontic and orthognathic treatments (Kwon et al., 2021). However, directly detecting cephalometric landmarks is a challenging task due to the inherent complexity of the highly non-linear mapping function (Pfister et al., 2015) that exists between the input cephalograms and the corresponding landmark locations. The intricate relationship between image features and landmark positions makes it difficult to achieve precise and reliable results.

We designed a two-stage regression framework to detect anatomical landmarks in cephalometric radiographs. The proposed framework is a multi-head CNN architecture that shares a single backbone neural network. The backbone network serves as a feature extractor, providing high-dimensional and semantically rich features to other modules. The landmark detection module extracts low-resolution but semantically strong features from the last layer of the backbone network to yield candidate region proposals for all anatomical landmarks. The detection head utilizes these features to simultaneously regress coordinates for all landmarks, enabling the framework to leverage both global hard/soft tissue characteristics and geometric landmark relations in a unified manner.

The performance of the landmark detection module resulted in an MRE value of $4.59 \pm 2.614$ mm, which necessitated the need for significant refinements to improve the accuracy of landmark predictions. To address this issue, we designed a refinement module that utilizes cropped patches of the feature map to fine-tune the coarsely predicted landmarks. Initially, our approach involved extracting the feature map from the last block of the backbone network and refining the predicted landmarks based on its cropped regions. However, this approach resulted in minimal improvements in landmark detection accuracy.

Upon further analysis, we encountered a significant challenge stemming from a substantial reduction in spatial dimensions (by a factor of 32) in the feature map of the last block. This reduction in dimensions became particularly pronounced since the original image had already been downscaled by a factor of 3. As a result, certain critical landmarks such as Menton, Gnathion, and Pogonion ended up being positioned in close proximity to one another and the region proposals surrounding these landmarks significantly overlapped (with an IoU exceeding 0.8) as illustrated in Fig. 7(a). Consequently, refinement module received nearly identical feature map regions for these three distinct landmarks, which hindered its ability to predict precise refinements.

In light of this challenge, we made the decision to extract feature regions from a block that would offer spatial dimensions substantial enough to prevent overlapping region proposals. However, such an approach would entail sacrificing the inclusion of global and semantically rich features, which could potentially compromise the performance of our refinement module. It is required to combine the high-resolution, semantically weak feature maps with the low-resolution, semantically rich ones to achieve a more refined landmark localization. Therefore, we introduce the semantic fusion block (SBF), which leverages feature maps from multiple blocks of the backbone network to produce high-resolution, semantically rich feature maps for accurate landmark refinement. This approach leads to rich semantics at all levels and can be quickly built from a single input image scale. The cropped high-resolution, semantically weak features are concatenated with low-resolution, semantically strong features, producing a single high-level feature map with fine resolution, which is then utilized for refining the landmark positions.

We conducted extensive experiments to evaluate the performance of our proposed framework against state-of-the-art methods on the ISBI 2015 dataset (Wang et al., 2015). In our initial experimental

**Table 2**
Comparison of our proposed framework with state-of-the-art methods in terms of mean radial error (MRE) ± standard deviation (SD) and successful detection rate (SDR) within four clinical ranges. The best results are highlighted in bold, while the second runner-up is underlined for a comprehensive comparison. The table clearly shows that our proposed framework outperformed state-of-the-art in the clinical ranges of 2.0 mm, 3.0 mm and 4.0 mm on Test-1, while achieving comparable results on Test-2.

| Research methodology | ISBI 2015 Test-1 dataset | | | | | ISBI 2015 Test-2 dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MRE ± SD (mm) | SDR (%) | | | | MRE ± SD (mm) | SDR (%) | | | |
| | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm |
| Ibragimov, Likar, Pernus, and Vrtovec (2015) | 1.84 ± 1.76 | 71.72 | 77.4 | 81.93 | 88.04 | – | 62.74 | 70.47 | 76.53 | 85.11 |
| Lindner and Cootes (2015) | 1.67 ± 1.65 | 74.95 | 80.28 | 84.56 | 89.68 | 1.92 ± 0.0 | 66.11 | 72.00 | 77.63 | 87.42 |
| Arik et al. (2017) | – | 75.37 | 80.91 | 84.32 | 88.25 | – | 67.68 | 74.16 | 79.11 | 84.63 |
| Lee et al. (2020) | 1.53 ± 1.74 | 82.11 | 88.63 | 92.28 | 95.96 | – | – | – | – | – |
| Zeng et al. (2021) | 1.34 ± 0.92 | 81.37 | 89.09 | 93.79 | 97.86 | 1.64 ± 0.91 | 70.58 | 79.53 | 86.05 | 93.32 |
| Qian et al. (2019) | 1.28 ± 0.0 | 82.51 | 86.25 | 89.31 | 90.62 | 1.54 ± 0.0 | 72.40 | 76.15 | 79.65 | 85.90 |
| Kwon et al. (2021) | **1.12 ± 0.0** | 86.91 | **91.44** | 94.21 | 97.68 | **1.41 ± 0.0** | **77.16** | **84.79** | **89.21** | **94.95** |
| **Ours** | 1.23 ± 0.89 | **87.17** | 90.77 | **95.18** | **98.12** | 1.78 ± 0.92 | 75.79 | 84.00 | 89.00 | 94.53 |

**Table 3**
Performance comparison of different backbone networks on ISBI 2015 Test-1 and Test-2 datasets. Notably, ResNet-50 (He, Zhang, Ren, & Sun, 2015a) demonstrates superior performance, exhibiting the lowest MRE of 1.51 ± 0.91 mm and the highest SDR values in the respective clinical ranges, namely 81.48%, 87.39%, 92.09%, and 96.32%.

| Backbones | ISBI 2015 Test-1 dataset | | | | | ISBI 2015 Test-2 dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MRE ± SD (mm) | SDR (%) | | | | MRE ± SD (mm) | SDR (%) | | | |
| | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm |
| VGG-16 | 2.17 ± 1.95 | 43.71 | 52.55 | 59.95 | 67.95 | 2.98 ± 1.87 | 34.79 | 43.95 | 51.86 | 62.58 |
| Darknet-19 | 1.97 ± 1.23 | 67.45 | 70.21 | 78.54 | 81.73 | 2.61 ± 2.18 | 47.32 | 56.93 | 67.25 | 72.03 |
| ResNet-34 | 1.59 ± 1.17 | 73.96 | 83.09 | 89.19 | 95.44 | 2.35 ± 2.20 | 56.26 | 68.32 | 76.37 | 87.11 |
| Darknet-53 | 1.27 ± 0.93 | 85.32 | 89.01 | 93.78 | 95.56 | 1.82 ± 1.03 | 75.11 | 82.93 | 85.38 | 89.79 |
| **ResNet-50** | **1.23 ± 0.89** | **87.17** | **90.77** | **95.18** | **98.12** | **1.78 ± 0.92** | **75.79** | **84** | **89** | **94.53** |

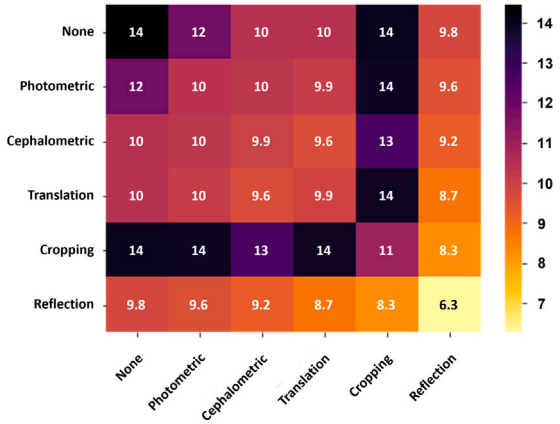



**Fig. 8.** Heatmap visualization illustrating the effects of individual and paired augmentation techniques on landmark detection accuracy, measured by mean radial error (mm).

phase, we aimed to finalize the training method (as discussed in Section 3.3) by contrasting the outcomes of joint training with those of separate training strategies. We observed that joint training produced results comparable to separate training while significantly reducing the training duration by approximately 25%–50%. Our results demonstrate significant improvements over existing approaches, highlighting the effectiveness of our solution. Specifically, our framework achieves an MRE of 1.23 ± 0.89 mm and the highest SDRs of 87.17%, 95.18% and 98.12% in clinical precision ranges of 2.0 mm, 3.0 mm and 4.0 mm, respectively.

### 6.1. Ablation studies

In this section, we present a comprehensive analysis of our proposed framework, assessing its performance and effectiveness from various perspectives. We begin by examining the impact of different backbone architectures on the framework's detection capabilities. Specifically, we evaluate the performance using a range of backbone networks, including VGG-16 (Simonyan & Zisserman, 2014), Darknet-19 (Redmon & Farhadi, 2016), ResNet-34 (He et al., 2015a), Darknet-53 (Redmon & Farhadi, 2018), and ResNet-50 (He et al., 2015a). Table 3 provides a detailed comparison of these networks on the ISBI 2015 Test-1 and Test-2 datasets. The results highlight ResNet-50 as the top performer, achieving the lowest MRE of 1.23 ± 0.89 mm and the highest SDRs across all clinical ranges: 87.17%, 90.77%, 95.18%, and 98.12%.

#### 6.1.1. Effects of augmentation techniques

We have conducted a comprehensive analysis to assess the effects of various augmentation techniques, as detailed in Section 4.2, on the overall objective function for landmark detection (refer to Eq. (4)). This analysis involved applying various augmentations individually and in pairs to the landmark detection module, utilizing VGG-16 (Simonyan & Zisserman, 2014) as the backbone. After training the model for 100 epochs and evaluating its performance on the testing dataset, we obtained quantitative results displayed as heatmap representation in Fig. 8. The results demonstrate the effectiveness of these augmentation techniques in improving landmark detection accuracy, quantified by mean radial error (MRE) in millimeters, reducing it from 14 mm (without augmentation) to 6.3 mm (with reflection). Notably, among the various techniques, reflection emerged as the most impactful, consistently improving accuracy when combined with other augmentation techniques. Conversely, cropping exhibited poor performance when paired with every other augmentation technique, leading to its exclusion from the augmented dataset.

**Table 4**
Performance of our proposed architecture with and without augmentation.

| Network configurations | MRE ± SD | SDR (%) | |
|---|---|---|---|
| | | 2.0 mm | 4.0 mm |
| LDM + LRM (without augmentation) | 3.339 ± 6.30 | 76.02 | 89.38 |
| LDM + LRM (with augmentation) | 1.853 ± 2.56 | 78.02 | 93.68 |

**Table 5**
Analysis of impact of different components of the proposed framework on landmark detection performance.

| Network | MRE ± SD (mm) | SDR (%) | | | |
|---|---|---|---|---|---|
| | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm |
| LDM only | 4.88 ± 2.56 | 10.67 | 14.00 | 24.67 | 41.33 |
| LDM + LRM (without SFB) | 1.91 ± 1.27 | 61.33 | 75.33 | 82.67 | 91.00 |
| LDM + LRM (with SFB) | 1.57 ±1.68 | 65.11 | 78.70 | 86.78 | 94.27 |

To further evaluate the performance of our proposed architecture under the influence of the assessed augmentation techniques, we conducted an additional experiment. We expanded the ISBI 2015 Train set by applying transformations identified as promising in minimizing the overall objective function. We trained the core architecture, comprising the landmark detection module (LDM) in conjunction with the landmark refinement module (LRM), coupled with the VGG16 backbone, with the augmented dataset for 150 epochs. Our observations, as illustrated in Table 4, revealed that in the absence of augmentation, the model tended to overfit the ISBI Train set due to its limited size. In contrast, with the augmented dataset, the model demonstrated generalized performance and learned finer anatomical features, enabling it to better discriminate between various craniofacial structures.

#### 6.1.2. Model component analysis

The results presented in Table 5 provide insights into the performance of different components of our proposed model for cephalometric landmark detection. Firstly, when considering only the landmark detection module (LDM), we observed a relatively high MRE of 4.88 ± 2.56 mm. This indicates that using only the LDM for landmark detection results in less accurate predictions, with a notable standard deviation suggesting inconsistency across different samples. The SDR values at various clinical precision thresholds are also relatively low, ranging from 10.67% to 41.33%. This indicates a higher percentage of landmarks falling outside the acceptable precision range, suggesting that relying solely on LDM leads to insufficient accuracy.

Next, we examined the performance of the combined Landmark Detection Module (LDM) and Landmark Refinement Module (LRM) without incorporating the Semantic Fusion Block (SFB). Here, we observed a notable improvement in MRE, which decreased to 1.91 mm ± 1.27 mm. This indicates that the inclusion of LRM helps refine the landmark predictions, resulting in more accurate localization. Correspondingly, the SDR values demonstrate substantial improvements across all precision thresholds compared to using the LDM alone. Moreover, incorporating the Semantic Fusion Block (SFB) along with the LRM further enhances the performance of the framework. The MRE decreased to 1.57 ± 1.68 mm, indicating even greater accuracy and precision in landmark detection. This underscores the importance of leveraging semantic information and multi-scale feature fusion provided by the SFB in refining landmark predictions and achieving higher accuracy.

#### 6.1.3. Model complexity analysis

We assess the complexity of our framework using two metrics: the number of parameters and the number of floating-point operations (FLOPs). Our framework comprises a landmark detection module (LDM), a semantic fusion block (SFB), and a landmark refinement module (LRM). Therefore, the total model complexity is the aggregate of these individual modules. As illustrated in Table 6, the total complexity of our framework amounts to approximately 457 million FLOPs and 38.7 million parameters. During the inference stage, we conducted runtime experiments on both GPU and a standard PC. As shown in 6, the inference process for a new image can be completed in about 3.37 s, which is considerably faster than the traditional manual approach used in clinical practice.

**Table 6**
Complexity and performance analysis of framework components.

| Network | Parameters (million) | FLOPs (million) | Execution time (s) | |
|---|---|---|---|---|
| | | | CPU | GPU |
| LDM | 9.7 | 9.7 | 0.53 | 0.07 |
| SFB | 2.1 | 309.3 | 1.83 | 1.05 |
| LRM | 10 | 138 | 1.01 | 0.79 |
| Total | 38.7 | 457 | 3.37 | 1.91 |

Additionally, comparing our proposed framework with the one by Zeng et al. (2021) who reported parameters of 69.11 million and FLOPs of 606.34 million, our model exhibits a remarkable improvement in efficiency. Specifically, our model is approximately 44% more efficient in terms of parameters and 24.6% more efficient in terms of FLOPs compared to Zeng et al. (2021). These findings indicate the superior efficiency and computational performance of our proposed framework compared to existing approaches.

### 6.2. Failure analysis

In our proposed framework, occlusions and misalignments of edge landmarks in cephalograms posed significant challenges. To investigate this further, we analyzed cases where landmarks were predicted with high mean radial errors. Fig. 9(a) illustrates an example of landmarks detection with an error of 4.915 ± 1.892 mm, exceeding the clinical accepted range of 2.0 mm. As depicted, the edge landmarks are partially obscured, making it difficult for the network to accurately identify them. This resulted in the generation of region proposals that were far from the ground truth, unable to provide sufficient regional context for the landmarks to be refined. Consequently, suboptimal refinements were made, leading to predictions that fell out of the clinical range.

Further examination of failure cases, depicted in Fig. 9(b), (c) and (d), highlighted a recurring issue: poor predictions by the landmark detection module leading to the generation of unreliable region proposals. These proposals lacked sufficient contextual information for effective refinement by the subsequent refinement heads, ultimately resulting in predictions falling outside the clinical range. To address this challenge, some graph-based repair strategies, as suggested by Qian et al. (2019), can be highly beneficial. This approach involves pre-processing the predictions obtained from the landmark detection module before region cropping, with the goal of improving the quality of the cropped regions by rectifying inaccurate predictions. Additionally, attention-based techniques, such as (Jiang et al., 2022; Lee, Chung, & Shin, 2022; Qian et al., 2020), could further improve the robustness of the landmark detection network and its ability to handle occlusions and misalignments effectively.

### 6.3. Extended experiments on PKU dataset

While the experiments on the ISBI 2015 dataset demonstrated the efficacy of our proposed framework, it is imperative to assess its generalization ability by subjecting it to other cephalometric landmark datasets. Therefore, we conducted a validation study of our proposed

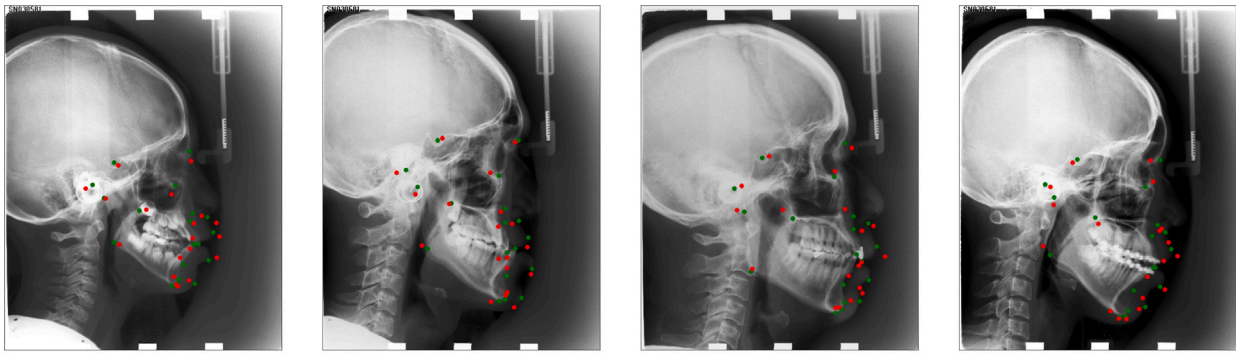



**Fig. 9.** An illustration of the cases where our proposed framework resulted in a high mean radial error (MRE) during landmark predictions. To demonstrate these cases, the predicted landmarks are sorted based on MRE, calculated with respect to the ground truth landmarks, and the top four cases are presented in order of increasing error from left to right.

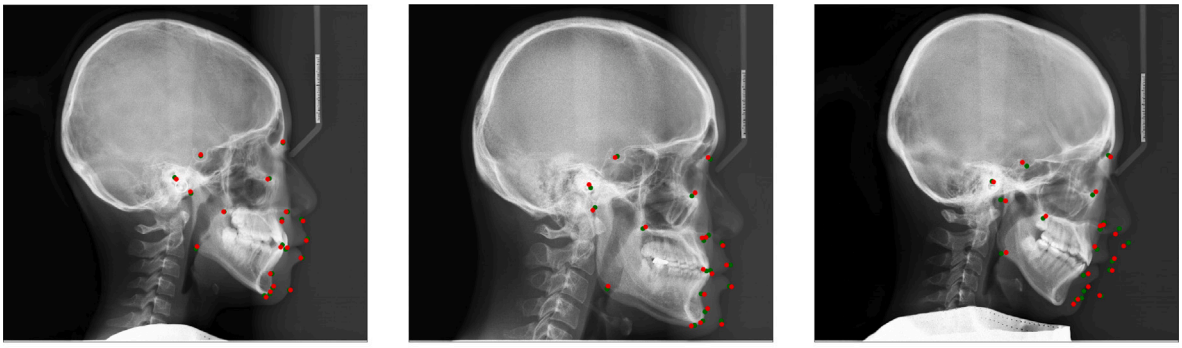



**Fig. 10.** Comparison of the performance of our proposed framework on the PKU dataset. (a) Illustrates the best results achieved, where the predicted landmarks (red) closely align with the ground truth landmarks (green). (b) Represents an intermediate level of performance, with some minor deviations between the predicted and ground truth landmarks. (c) Illustrates the worst-case scenario, where the predicted landmarks exhibit significant deviations from the ground truth landmarks.

framework on the PKU cephalogram dataset,[3] which was introduced by Zeng et al. (2021). This dataset includes 102 cephalograms with an average spatial dimension of $2089 \times 1937 \times 3$ pixels and a resolution of approximately 0.125 mm/pixel. Each image was independently annotated by two expert orthodontists with 19 cephalometric landmarks. The landmark prediction results are summarized in Table 7. The results demonstrate that our proposed framework can predict landmarks within clinically accepted range of 2 mm, even without fine-tuning on the new dataset.

In addition, we further analyzed the prediction results and show the best, average, and worst cases sorted by mean radial error (MRE) in Fig. 10 from left to right. Notably, the worst case (shown in Fig. 10(c)) demonstrates a consistent offset in the predicted landmarks compared to the ground truth. The reason for this failure is likely due to the head scale in this cephalogram being significantly different from the other cases, which was not learned precisely by the backbone network. Moreover, we observed that the MRE for the PKU dataset is higher than that for the ISBI 2015 dataset. This is mainly because the PKU dataset contains images with different resolutions and varying head sizes, leading to more challenging landmark localization. Despite these challenges, the proposed framework still achieved promising results, demonstrating its generalization ability across different datasets and can be applied to clinical practice without the need for fine-tuning.

[3] https://doi.org/10.6084/m9.figshare.13265471.v1

**Table 7**
Comparison of the performance of our proposed framework in terms of mean radial error (MRE) ± standard deviation (SD) and success detection rate (SDR) on the PKU cephalometric landmark dataset.

| Research methodology | MRE ± SD (mm) | SDR (%) | | | |
|---|---|---|---|---|---|
| | | 2.0 mm | 2.5 mm | 3.0 mm | 4.0 mm |
| Zeng et al. (2021) | 2.02 ± 1.89 | 64.81 | 73.94 | 81.73 | **89.78** |
| **ours** | **1.91 ± 1.78** | **67.34** | **75.18** | **83.49** | 89.63 |

### 6.4. Clinical relevance

In this section, we aim to discuss the clinical relevance and applicability of our proposed framework. To evaluate the clinical usability of our framework, we conducted a real-world evaluation in a clinical setting. We presented 100 cephalograms to both an experienced orthodontist and our framework, and calculated the agreement between both entities. Our framework demonstrated promising results, with 61.33% of the landmarks detected within a 2 mm range. However, we observed a relatively high mean radial error of approximately 4.051 ± 2.409. One plausible explanation for this discrepancy is that our framework was trained on a dataset comprising cephalograms captured at a specific resolution by a single imaging device, potentially introducing bias. Therefore, there is a compelling need for a dataset containing multi-resolution cephalograms obtained from various imaging devices to enhance the robustness and generalizability of our framework.

Recently, the Aariz dataset (Khalid et al., 2022, 2023) marks a pivotal development in the field. This dataset, containing 1000 multi-resolution cephalograms sourced from 7 distinct imaging devices, presents a compelling opportunity for advancing our research. Although our current framework demonstrates efficiency and robustness, we anticipate that extending our study to include datasets like Aariz will enhance our framework's ability to generalize. Such an expansion aligns with our goal of integrating our framework into routine clinical practices, where variability in imaging devices and resolutions is the norm.

### 6.5. Current challenges and future directions

The results presented in Section 5 clearly demonstrate the significant contribution made by our proposed framework towards automated cephalometric analysis, achieving state-of-the-art level results on the IEEE ISBI 2015 dataset by Wang et al. (2015). However, despite this achievement, there remain several challenges that require further attention.

One of the primary challenges in developing an accurate cephalometric landmark detection model is the limited size of the available training dataset. In the case of the IEEE ISBI 2015 Dataset, the training data consists of only 150 cephalograms obtained randomly selected from a pool of 400 patients ranging in age from six to 60 years old. This limited sample size and diverse patient pool can make it difficult for an AI algorithm to effectively generalize and may lead to overfitting (Domingos, 2012). Moreover, the inter- and intra-observer variability between the two orthodontists who labeled the dataset introduces an additional layer of uncertainty. Specifically, the mean intra-observer variability for the senior and junior orthodontists is $1.73 \pm 1.35$ mm and $0.90 \pm 0.89$ mm, respectively, while the mean inter-observer variability is $1.38 \pm 1.55$ mm, resulting in a mean radial error (MRE) of $2.02 \pm 1.53$ mm on test data. Given the clinical precision range of 2 mm, this degree of variability is significant and may result in unnecessary bias being introduced into the trained model (Lee et al., 2020). Therefore, there is a need for new state-of-the-art datasets that can help overcome these challenges. The recently introduced Aariz dataset (Khalid et al., 2022, 2023), as discussed in Section 6.4, marks a pivotal development in this regard. This dataset presents a compelling opportunity for advancing research by providing a more diverse and extensive set of images, which can enhance the generalizability of AI algorithms. Expanding our study to include datasets like Aariz will likely improve our framework's robustness and its applicability to clinical practice, where variability in imaging devices and resolutions is common.

Similarly, CNNs have made significant efforts to learn the graphical structure of landmarks, but they still fall short in capturing the intrinsic geometrical relations between the landmarks. The landmarks in medical images are not located independently rather their positioning is constrained by the presence of other landmarks, and this structural knowledge has not been fully utilized in existing methods. In this regard, graph convolutional networks may prove to be a promising solution, as many researchers (Li et al., 2020; Lu et al., 2022) have recently explored this direction. However, further exploration is necessary to fully unlock the potential.

## 7. Conclusion

This research addresses the crucial need for accurate cephalometric landmark detection in orthodontics, offering a promising solution to automate the tracing process and enhance clinical efficiency. By introducing a two-stage regression framework with a shared backbone neural network, we have significantly overcome the limitations of existing approaches, which rely on a one-to-one mapping between the landmarks and the CNNs. Through the incorporation of high-dimensional and semantically rich features, along with the simultaneous regression of coordinates for all landmarks, our framework leverages both global hard/soft tissue characteristics and geometric landmark relations in a unified manner, providing a comprehensive understanding of craniofacial deformities. The results obtained from our framework showcase its potential to revolutionize cephalometric analysis by reducing subjectivity and the time required for manual landmark identification. We anticipate that our proposed framework will assist in improving patient outcomes, advancing treatment strategies, and facilitating comprehensive assessment of craniofacial abnormalities.

## CRediT authorship contribution statement

**Muhammad Anwaar Khalid:** Conceptualization, Methodology, Software, Experimentation, Writing – original draft. **Atif Khurshid:** Conceptualization, Experimentation. **Kanwal Zulfiqar:** Data curation. **Ulfat Bashir:** Data curation, Conceptualization. **Muhammad Moazam Fraz:** Conceptualization, Writing – original draft, Writing – review & editing, Project supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Code is available online. Link in the manuscript.

## Acknowledgments

We acknowledge the support of the Islamic World Educational, Scientific, and Cultural Organization (ICESCO) in establishing the ICESCO Chair of Data Science and Analytics for Business at the National University of Sciences and Technology (NUST). This initiative has made a significant contribution to the advancement of research and academic pursuits in applied AI.

## References

Arik, S. Ö., Ibragimov, B., & Xing, L. (2017). Fully automated quantitative cephalometry using convolutional neural networks. *Journal of Medical Imaging*, *4*(1), Article 014501.

Bulat, A., & Tzimiropoulos, G. (2018). Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 109–117).

Cardillo, J., & Sid-Ahmed, M. A. (1994). An image processing system for locating craniofacial landmarks. *IEEE Transactions on Medical Imaging*, *13*(2), 275–289.

Chen, R., Ma, Y., Chen, N., Lee, D., & Wang, W. (2019). 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, vol. 22* (pp. 873–881). Springer.

Chen, C., & Zheng, G. (2014). Fully-automatic landmark detection in cephalometric x-ray images by data-driven image displacement estimation. In *Proceedings of international symposium on biomedical imaging* (pp. 17–24).

Chu, C., Chen, C., Nolte, L.-P., & Zheng, G. (2014). Fully automatic cephalometric x-ray landmark detection using random forest regression and sparse shape composition. In *Proceedings of international symposium on biomedical imaging* (pp. 9–16).

Deng, G. (2010). A generalized unsharp masking algorithm. *IEEE Transactions on Image Processing*, *20*(5), 1249–1261.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

El-Feghi, I., Sid-Ahmed, M. A., & Ahmadi, M. (2004). Automatic localization of craniofacial landmarks for assisted cephalometry. *Pattern Recognition*, *37*(3), 609–621.

Forsyth, D., & Davis, D. (1996). Assessment of an automated cephalometric analysis system. *European Journal of Orthodontics*, *18*(5), 471–478.

Grau, V., Alcaniz, M., Juan, M., Monserrat, C., & Knoll, C. (2001). Automatic localization of cephalometric landmarks. *Journal of Biomedical Informatics*, *34*(3), 146–156.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). Mask R-CNN. arXiv:1703.06870.
He, K., Zhang, X., Ren, S., & Sun, J. (2015a). Deep residual learning for image recognition. arXiv:1512.03385.
He, K., Zhang, X., Ren, S., & Sun, J. (2015b). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. arXiv:1502.01852.
Hou, Q., Wang, J., Cheng, L., & Gong, Y. (2015). Facial landmark detection via cascade multi-channel convolutional neural network. In *2015 IEEE international conference on image processing* (pp. 1800–1804). IEEE.
Ibragimov, B., Likar, B., Pernuš, F., & Vrtovec, T. (2014). Automatic cephalometric X-ray landmark detection by applying game theory and random forests. In *Proceedings of international symposium on biomedical imaging* (pp. 1–8).
Ibragimov, B., Likar, B., Pernus, F., & Vrtovec, T. (2015). Computerized cephalometry by game theory with shape-and appearance-based landmark refinement. In *Proceedings of international symposium on biomedical imaging*.
Jiang, Y., Li, Y., Wang, X., Tao, Y., Lin, J., & Lin, H. (2022). CephalFormer: incorporating global structure constraint into visual features for general cephalometric landmark detection. In *International conference on medical image computing and computer-assisted intervention* (pp. 227–237). Springer.
Kamoen, A., Dermaut, L., & Verbeeck, R. (2001). The clinical significance of error measurement in the interpretation of treatment results. *The European Journal of Orthodontics*, *23*(5), 569–578.
Khalid, M. A., Zulfiqar, K., Bashir, U., Shaheen, A., Iqbal, R., Rizwan, Z., et al. (2022). CEPHA29: Automatic cephalometric landmark detection challenge 2023. arXiv preprint arXiv:2212.04808.
Khalid, M. A., Zulfiqar, K., Bashir, U., Shaheen, A., Iqbal, R., Rizwan, Z., et al. (2023). 'Aariz: A benchmark dataset for automatic cephalometric landmark detection and CVM stage classification. arXiv preprint arXiv:2302.07797.
Koonsanit, K., Thongvigitmanee, S., Pongnapang, N., & Thajchayapong, P. (2017). Image enhancement on digital x-ray images using N-CLAHE. In *2017 10th biomedical engineering international conference* (pp. 1–4). IEEE.
Kwon, H. J., Koo, H. I., Park, J., & Cho, N. I. (2021). Multistage probabilistic approach for the localization of cephalometric landmarks. *IEEE Access*, *9*, 21306–21314.
Lee, M., Chung, M., & Shin, Y. G. (2022). Cephalometric landmark detection via global and local encoders and patch-wise attentions. *Neurocomputing*, *470*, 182–189.
Lee, H., Park, M., & Kim, J. (2017). Cephalometric landmark detection in dental x-ray images using convolutional neural networks. *vol. 10134*, In *Medical imaging 2017: computer-aided diagnosis* (pp. 494–499). SPIE.
Lee, J. H., Yu, H. J., Kim, M. j., Kim, J. W., & Choi, J. (2020). Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC Oral Health*, *20*(1), 1–10.
Li, W., Lu, Y., Zheng, K., Liao, H., Lin, C., Luo, J., et al. (2020). Structured landmark detection via topology-adapting deep graph learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part IX, vol. 16* (pp. 266–283). Springer.
Lindner, C., & Cootes, T. F. (2015). Fully automatic cephalometric evaluation using random forest regression-voting. In *IEEE international symposium on biomedical imaging*. Citeseer.
Lu, G., Zhang, Y., Kong, Y., Zhang, C., Coatrieux, J. L., & Shu, H. (2022). Landmark localization for cephalometric analysis using multiscale image patch-based graph convolutional networks. *IEEE Journal of Biomedical and Health Informatics*, *26*(7), 3015–3024.
Maini, R., & Aggarwal, H. (2010). A comprehensive review of image enhancement techniques. arXiv preprint arXiv:1003.4053.
Milošević, D., Vodanović, M., Galić, I., & Subašić, M. (2022). Automated estimation of chronological age from panoramic dental X-ray images using deep learning. *Expert Systems with Applications*, *189*, Article 116038. http://dx.doi.org/10.1016/j.eswa.2021.116038, URL https://www.sciencedirect.com/science/article/pii/S095741742101383X.
Mirzaalian, H., & Hamarneh, G. (2014). Automatic globally-optimal pictorial structures with random decision forest based likelihoods for cephalometric x-ray landmark detection. In *Proceedings of international symposium on biomedical imaging* (pp. 25–36).
Mohseni, H., & Kasaei, S. (2007). Automatic localization of cephalometric landmarks. In *2007 IEEE international symposium on signal processing and information technology* (pp. 396–401). IEEE.
Oh, K., Oh, I. S., Lee, D. W., et al. (2020). Deep anatomical context feature learning for cephalometric landmark detection. *IEEE Journal of Biomedical and Health Informatics*, *25*(3), 806–817.
Pfister, T., Charles, J., & Zisserman, A. (2015). Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 1913–1921).
Qian, J., Cheng, M., Tao, Y., Lin, J., & Lin, H. (2019). CephaNet: An improved faster R-CNN for cephalometric landmark detection. In *2019 IEEE 16th international symposium on biomedical imaging* (pp. 868–871). IEEE.
Qian, J., Luo, W., Cheng, M., Tao, Y., Lin, J., & Lin, H. (2020). CephaNN: a multi-head attention network for cephalometric landmark detection. *IEEE Access*, *8*, 112633–112641.
Redmon, J., & Farhadi, A. (2016). YOLO9000: Better, faster, stronger. arXiv:1612.08242.
Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767.
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.
Ruder, S. (2017). An overview of gradient descent optimization algorithms. arXiv:1609.04747.
Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 1–48.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Singh, N. K., & Raza, K. (2022). Progress in deep learning-based dental and maxillofacial image analysis: A systematic review. *Expert Systems with Applications*, *199*, Article 116968.
Sun, Y., Wang, X., & Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3476–3483).
Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in Neural Information Processing Systems*, *27*.
Vandaele, R., Marée, R., Jodogne, S., & Geurts, P. (2014). Automatic cephalometric x-ray landmark detection challenge 2014: a tree-based algorithm. In *Proceedings of international symposium on biomedical imaging* (pp. 37–44).
Wang, C. W., Huang, C. T., Hsieh, M. C., Li, C. H., Chang, S. W., Li, W. C., et al. (2015). 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. http://dx.doi.org/10.1109/TMI.2015.2412951.
Wang, C. W., Huang, C. T., Lee, J. H., Li, C. H., Chang, S. W., Siao, M. J., et al. (2016). A benchmark for comparison of dental radiography analysis algorithms. *Medical Image Analysis*, *31*, 63–76.
Yue, W., Yin, D., Li, C., Wang, G., & Xu, T. (2006). Automated 2-D cephalometric analysis on X-ray images by a model-based approach. *IEEE Transactions on Biomedical Engineering*, *53*(8), 1615–1623.
Zeng, M., Yan, Z., Liu, S., Zhou, Y., & Qiu, L. (2021). Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*, *68*, Article 101904.
Zhu, M., Shi, D., & Gao, J. (2019). Branched convolutional neural networks incorporated with Jacobian deep regression for facial landmark detection. *Neural Networks*, *118*, 127–139.