Contour Makes It Stronger: Cross-Domain Cephalometric Landmark Detection Based on Contour Priors

Xinyue Liang[1], Runnan Chen[2], Guangshun Wei[1]*, Shaojie Zhuang[1], and Yuanfeng Zhou[1]*

[1] School of Software, Shandong University, Jinan, China
yfzhou@sdu.edu.cn
[2] Sydney Artificial Intelligence Centre, the University of Sydney, Sydney, Australia

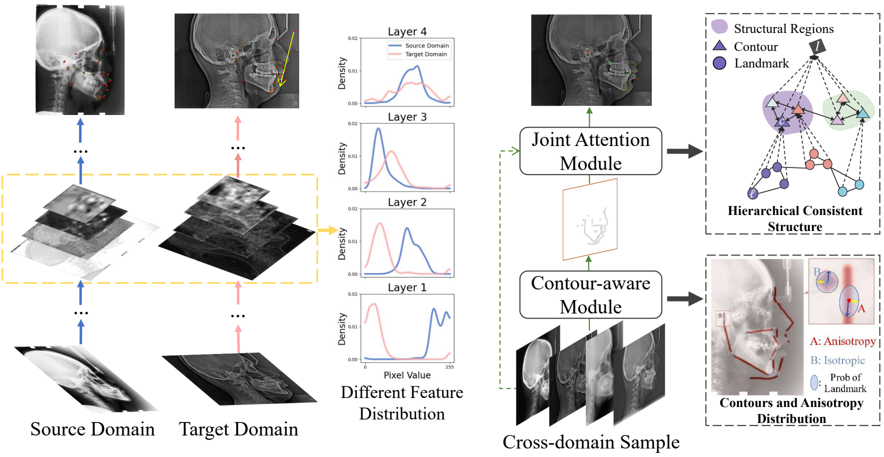
Abstract. The detection of cephalometric landmarks is crucial for orthodontic diagnosis. Current methods mainly focus on utilizing contextual information to detect landmarks while overlooking the challenges posed by domain gaps. In this paper, we propose a contour-guided framework that leverages cranial soft/hard tissue contours as domain-invariant anatomical priors. The method introduces a joint attention module to fuse the topological features corresponding to the contours with contextual features, ensuring the accuracy of landmark positioning. Additionally, we address anisotropic prediction uncertainty in unseen domains through a direction-aware regression module, which incorporates contour geometry to regularize error distributions. Evaluated on the multi-domain datasets with five source and three unseen target domains, our framework demonstrates superior robustness to domain shifts while maintaining anatomical plausibility, achieving state-of-the-art cross-domain localization accuracy.

Keywords: Cephalometric Landmark · Domain Gap · Feature Fusion.

1 Introduction

Cephalometric analysis utilizes 2D images generated by X-ray, providing clinicians with crucial information on patients' dental, skeletal, and facial relationships. As an indispensable part of orthodontic and orthognathic treatments, a pivotal step in this process is the detection of key anatomical landmarks. In practice, landmarks are located manually, which is tedious, time-consuming, and unreliable in achieving reproducible results. Therefore, fully automatic and accurate landmark localization has been a long-standing area of significant need.

Recent deep learning-based landmark detection methods [1, 3, 12] have advanced through multi-scale contextual information extraction and global-local feature interaction in single domain application. For instance, Chen et al. [3] developed an attentive feature pyramid fusion module for feature fusing, while Lee



(a) Existing methods utilize only landmark-based features(sensitive to image appearance).

(b) Our method based on contour-landmark joint features shows robustness across domains.

Fig. 1: Difference between existing methods and ours.

et al. [11] modeled spatial structure via global coordinate regression and local feature cropping. However, domain gap caused by variations in image properties (e.g., resolution, contrast) significantly impair cross-domain generalization, as shown in Fig. 1(a). The yellow rectangle highlights the differences in feature distributions across domains, particularly in lower layers, which are highly sensitive to image appearance variations.

To address this cross-domain degradation, researchers have employed strategies like domain adaptation. Jin et al. [8] combined self-training with adversarial learning for knowledge transfer, and Wu et al. [16] enhanced robustness through anatomical prototype relation mining. Despite reducing annotation dependency, these methods still require retraining on the target domain, limiting clinical scalability due to data acquisition costs and workload.

Compared to domain adaptation, we adopt domain generalization by extracting domain-invariant features, as cranial soft/hard tissue contours, to bridge domain gaps in cephalometric landmark detection. Unlike landmarks, which are sensitive to image appearance, these contours encode continuous geometric priors that inherently resolve domain gaps. As shown in Fig. 1(b), by learning associations between landmarks and anatomically defined boundaries, our method establishes task-specific structural constraints, forcing the network to focus on domain-agnostic topology.

In this paper, we propose a contour-aware joint learning (CJL) framework for cross-domain cephalometric landmark detection. The method first models cranial soft/hard tissue contours as domain-invariant priors to capture structural consistency. A multi-scale CNN extracts contextual features, which are then fused with contour features as structural features through a joint attention

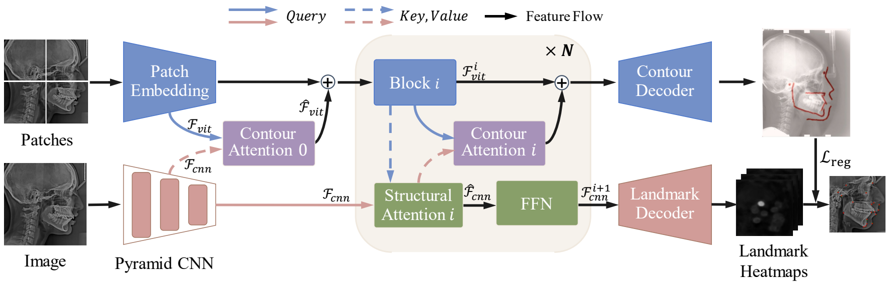


Fig. 2: The pipeline of our method: the top branch, represented by several blue blocks, corresponds to the Contour-aware Module, the shaded area denotes the Joint Attention Module, and L_{reg} indicates the Direction-sensitive Regression Module. The Pyramid CNN encodes contextual features, while the Landmark Decoder generates landmark heatmaps.

module (JAM). This module aligns contour features and landmark features in a shared embedding space, generating globally consistent hierarchical features. Furthermore, we observed that landmark predictions in the target domain exhibit higher uncertainty along the tangential direction of contours compared to the normal direction. This anisotropic distribution motivated us to propose a direction-sensitive regression module (DRM). By guiding landmark regression based on the tangential and normal directions of contours, our method incorporates anatomical plausibility into landmark prediction, enhancing robustness. The contributions of this paper are as follows:

1. We explore domain generalization by using cranial contours as domain-invariant priors for cephalometric landmark detection.
2. We propose a contour-landmark joint attention module to generate globally consistent hierarchical features for cross-domain generalization.
3. We introduce a novel regression strategy that leverages contour structure to mitigate anisotropic prediction uncertainty, enhancing cross-domain stability.
4. Our method achieves state-of-the-art performance on unseen target domains, demonstrating superior generalization capability.

2 Method

Domain gaps present a significant barrier in landmark detection tasks, particularly when dealing with variations in imaging devices and parameters across clinical environments. Our contour-aware joint learning (CJL) framework addresses this challenge by modeling cranial contours as anatomical structural priors that exhibit intrinsic invariance across domains. As depicted in Fig. 2, the framework consists of three synergistic components: the Contour-aware Module (Sec.2.1), the Joint Attention Module (Sec.2.2), and the Direction-sensitive Regression Module(Sec 2.3).

2.1 Contour-aware Module

The inherent invariance of anatomical contours across different domains stems from the structural consistency of cranial anatomy. Specifically, contours serve as stable structural priors,reducing reliance on low-level features (e.g., pixel intensity gradients) that are prone to scanner-induced artifacts. Moreover, they enforce anatomical plausibility through explicit structural constraints (Delaire's analysis [4]). Compared to direct landmark detection, this hierarchical representation establishes an interpretable mapping that aligns with clinical reasoning processes.

Medical-defined landmarks $\mathbf{P} = \left\{p_i \in \mathbb{R}^2\right\}_{i=1}^{L}$ are interpolated into $\mathbf{N}$ anatomical contours using cubic splines:

$$c_j(t) = \sum_{k=0}^{3} \mathbf{B}_k(t) \cdot p_{m+k}, \quad t \in [0,1], \tag{1}$$

where $\mathbf{B}_k(t)$ denotes basis functions, and p_{m+k} are the landmarks belonging to contour c_j, with m as the starting index of the landmarks, and $k \in \{0,1,2,3\}$, indicating the use of four control points. Then, for each contour c_j, we compute a distance transform D_j and Gaussian-smoothed heatmap M_j [18]:

$$D_j(x,y) = \min_{(x',y') \in c_j} |(x,y) - (x',y')|_2, \tag{2}$$

$$M_j(x,y) = \exp\left(-\frac{D_j(x,y)^2}{2\sigma^2}\right) \cdot \mathbb{I}(|D_j(x,y)| \leq 3\sigma), \tag{3}$$

where σ controls the spatial uncertainty, and we set $\sigma = 1.0$. Here, $\mathbb{I}(\cdot)$ is an indicator function.

To obtain geometric features and sematic anatomical information, we employ a pretrained Vision Transformer (ViT) as the image encoder [10]. Given a cephalometric image I, it is first cropped into non-overlapping 16 × 16 patches, which are then flattened and projected into D-dimensional tokens. After adding position embeddings, these tokens are fed into M multi-head self-attention (MSA) layers and MLP blocks [5], ultimately generating a contour feature map with a resolution of $\frac{1}{16}$ of the original image.

2.2 Joint Attention Module

The hierarchical topology between contours and landmarks establishes constraints for feature fusion. As depicted in Fig. 1(b), the Hierarchical Consistent Structure comprises three structural hierarchies: base-level landmarks l, intermediate contours c , and overlying tissue regions (shaded areas), encapsulating the structural topology with both intra-layer and cross-layer relationships. To capture these relationships, we develop a joint attention mechanism that enables cross-hierarchy message passing through alternating query strategies.

Given the structural features $F_{vit} \in \mathbb{R}^{(\frac{H}{16} \times \frac{W}{16}) \times D}$ and the contextual features $F_{cnn} = \{F_k\}_{k=1}^{3}$with resolutions of $\frac{1}{8}, \frac{1}{16}, \frac{1}{32}$, we iteratively apply Eq. (4) and Eq. (5) for N rounds. In each round, we take F^i_{vit} as the query, and F^i_{cnn} as the key and value for the contour attention layer, as shown in Eq. (4). The resulting feature $\hat{\mathcal{F}}^i_{\text{vit}}$ is then passed through the encoder layer in the i-th block to produce the updated feature $\mathcal{F}^{i+1}_{vit}$. After that, F^i_{cnn} is used as the query, and F^{i+1}_{vit} as the key and value for the structural attention layer, as shown in Eq. (5).

$$\hat{\mathcal{F}}^i_{vit} = \mathcal{F}^i_{vit} + \gamma^i \, \text{Attention}\left(\text{norm}\left(\mathcal{F}^i_{vit}\right), \text{norm}\left(\mathcal{F}^i_{cnn}\right)\right), \tag{4}$$

$$\hat{\mathcal{F}}^i_{\text{cnn}} = \mathcal{F}^i_{\text{cnn}} + \text{Attention}\left(\text{norm}\left(\mathcal{F}^i_{\text{cnn}}\right), \text{norm}\left(\mathcal{F}^{i+1}_{\text{vit}}\right)\right), \tag{5}$$

where the $norm(\cdot)$ is LayerNorm [2] and the attention layer is Attention$(\cdot)$ suggests using sparse attention [20].

2.3 Direction-sensitive Regression Module

In cross-domain landmark detection, the prediction distributions of landmarks exhibit relevance to contours. As visualized in Fig. 1(b) ("Contours and Anisotropy Distribution"), arrows encode the deviation of landmark predictions. For landmark A on the contour, its probability distribution exhibits higher variance along the contour's tangential direction $\mathit{\Delta} t$ and lower variance along the normal direction $\mathit{\Delta} n$, demonstrating anisotropic characteristics (blue ellipse). In contrast, non-contour landmark B shows isotropic uncertainty with uniform variance around the ground truth. Building upon this correlation, we propose a contour-aware regression strategy that adjusts error tolerance along the tangential $\mathbf{t}$ and normal $\mathbf{n}$ directions.

From the contour heatmaps $F_{contour}$ generated by the contour-aware module, we generate contour-aware direction fields by computing pixel-wise tangent $\mathbf{t}$ and normal $\mathbf{n}$ vectors via:

$$\mathbf{t}(x, y) = \frac{\left(\frac{\partial M_j}{\partial y}, -\frac{\partial M_j}{\partial x}\right)}{|\nabla F_{contour}|_2}, \mathbf{n}(x, y) \quad = \frac{\left(\frac{\partial M_j}{\partial x}, \frac{\partial M_j}{\partial y}\right)}{|\nabla F_{contour}|_2}, \tag{6}$$

where partial derivatives are computed using Sobel operators, ensuring differentiability. We then project the offset between the predicted landmark p and its ground truth p^* onto the directional vectors, obtaining the tangent $\mathit{\Delta} t$ and normal $\mathit{\Delta} n$ offset components.

To adaptively address the anisotropic distribution, we compute direction-specific energy values as moving averages of squared errors:

$$E^k_t = \alpha E^{k-1}_t + (1 - \alpha)\mathit{\Delta} t^2, \quad E^k_n = \alpha E^{k-1}_n + (1 - \alpha)\mathit{\Delta} n^2, \tag{7}$$

with a smoothing factor $\alpha = 0.9$. The regression loss becomes:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left(\frac{\Delta t^i}{E_t + \epsilon} + \frac{\Delta n^i}{E_n + \epsilon} \right), \quad (8)$$

where ϵ is set to 10^{-5} to prevent division by zero.

In the prediction module, we combine the contour heatmaps, landmark heatmaps and anisotropic offsets to predict landmark positions. The loss function L_h is defined to be mean logistic losses between the predicted landmark heatmaps and the ground truth. The loss function L_{mse} is defined to be the L2 loss between the predicted contour heatmaps and the ground truth heatmap M_j. The final loss fuction is defined as follows:

$$\mathcal{L} = \mathcal{L}_{reg} + \lambda_1 \mathcal{L}_h + \lambda_2 \mathcal{L}_{mse}, \quad (9)$$

where $\lambda_1 = 2.0, \lambda_2 = 5.0$.

3 Experiments

3.1 Experimental Settings

Dataset: To validate cross-domain generalization, we construct a multi-domain benchmark combining the ISBI 2023 Challenge dataset [9] (700 images from 7 different devices including Planmeca ProMax® 3D, Hyperion X5 2D PAN CEPH and so on) and ISBI 2015 dataset [15] (400 images). For cross-domain setting, we split the data into five source domains (832 training images) and three unseen target domains (400 test images) based on acquisition devices, with resolution varying from 1280 × 960 to 2560 × 1920. All evaluations adopt the standardized 19-landmark definitions from ISBI 2015.

Contours: Eight contours are involved in this work (e.g., maxillary bone outline, mandibular bone outline). These contours are well-established in cephalometric literature [7] and have been validated by clinical experts. The contour structures and landmark-contour mapping rules (e.g., Points Pogonion, Meton, Gonion on the mandibular bone outline) remain consistent and valid even with an increasing number of landmarks.

Implementation Details: We employ the ImageNet-1K pre-trained weights from DeiT (Touvron et al. [14]) to initialize our ViT-B model with embedding dimensions of 768, depth of 12, and 12 attention heads. A pre-trained VGG-19 [13] is as the multi-scale CNN backbone. To save time and memory, we set up a joint learning interaction for every 4 embedding layers, meaning the JAM runs M=3 times. The entire framework is optimized using the Adadelta optimizer with default configurations. The training process takes approximately 4 hours for 150 epochs on three Geforce RTX 3090 GPUs.

Evaluation Metrics: Following previous studies [15], we evaluate the model's performance using two commonly used metrics: 1) Mean Radial Error (MRE), which calculates the average Euclidean distance between the predicted and

Table 1: Results on source domain and target domain, respectively.

Model	Source Domain Test Dataset					Target Domain Test Dataset				
	MRE ↓	SDR(%) ↑				MRE ↓	SDR(%) ↑			
	(mm)	1mm	2mm	3mm	4mm	(mm)	1mm	2mm	3mm	4mm
FPA [6]	1.89	29.56	63.58	82.27	89.92	4.43	4.58	15.10	30.21	46.56
YOLOs [19]	1.27	55.42	82.01	91.47	95076	2.41	44.98	68.56	82.05	88.32
AFPF [3]	1.13	58.05	**86.87**	**95.22**	**97.96**	1.81	47.34	76.62	83.61	83.58
Wu [17]	1.12	62.88	86.04	93.85	96.72	1.62	51.35	75.89	86.09	91.41
CeLDA [16]	1.15	56.24	84.26	94.33	97.94	1.51	54.07	80.40	88.33	92.07
Ours	**1.09**	**64.15**	86.55	93.95	97.10	**1.43**	**55.25**	**80.64**	**89.58**	**93.78**

ground-truth landmarks; and 2) Successful Detection Rate (SDR), defined as the percentage of landmarks accurately detected within distances of 1.0mm, 2.0mm, 3.0mm, and 4.0mm from the ground-truth landmarks. It is worth mentioning that we specially compared SDR within an extremely small error range (within 1 mm) to validate the accuracy of landmark detection under high precision.

3.2 Comparison with SOTA Approaches

We compare our CJL with several state-of-the-art cephalometric landmark detection models, including recent classical methods (AFPF [3], FPA [6], YOLO [19]), and CelDA [16]—a method specifically designed to address the domain gap between adults and adolescents. We also compare with the recent champion method proposed by Wu et al. [17]. To ensure fairness, all competing approaches were retrained with the same configurations on our cross-domain dataset.

From Table 1, we can observe that other models perform significantly worse on the target domain compared to the source domain. For example, AFPF shows a higher MRE (1.81 vs. 1.13) and lower SDR (76.62% vs. 86.04% within 2mm), indicating that the domain gap leads to severe performance degradation. Although our method does not achieve optimal values across all source domain metrics, it consistently outperforms other competing approaches in all metrics. Notably, our model exhibits only a 0.34mm MRE degradation compared to source domain performance, demonstrating superior cross-domain robustness compared to existing approaches (CeLDA: 0.36mm; Wu et al.: 0.5mm; AFPF: 0.68mm; YOLO: 1.14mm; FPA: 2.54mm).

As shown in Fig. 3, we present the comparative performance across methods under significant appearance discrepancies between source and target domain datasets. On the target domain, AFPF and YOLO exhibit substantial landmark displacements (indicated by yellow arrows). In contrast, while other methods avoid such failures, our approach demonstrates enhanced precision compared to Wu and CeLDA. The yellow rectangles highlight CJL's superior localization accuracy in braces-present/absent cases. The blue rectangles further validate our method's enhanced precision in contour lines.

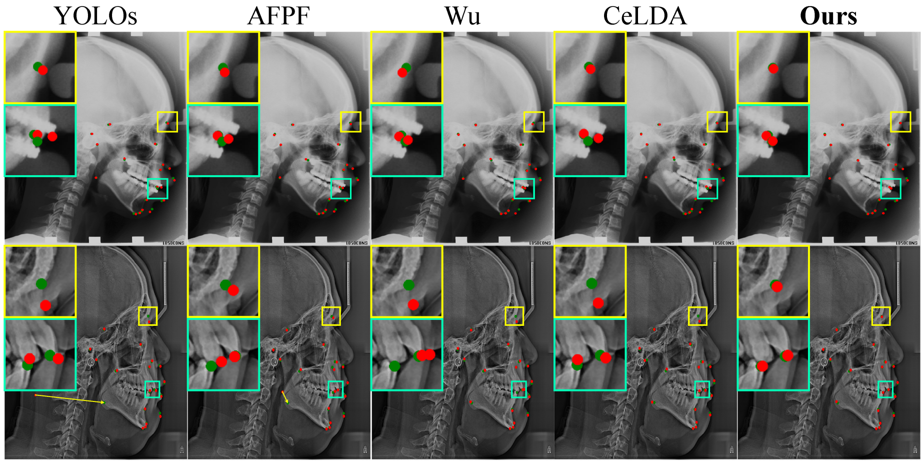


Fig. 3: Qualitative results of three models on target domain test data. Green dots are GTs, and red dots are predictions. Yellow arrows denote significant displacement between predictions and GTs. Rectangles indicate that our model performs better than the others.

Table 2: Ablation analysis for our proposed CJL method.

			Source Domain					Target Domain				
CAM	JAM	DRM	MRE ↓	SDR(%) ↑				MRE ↓	SDR(%) ↑			
			(mm)	1mm	2mm	2.5mm	3mm	(mm)	1mm	2mm	2.5mm	3mm
			1.44	59.84	78.95	90.05	92.79	1.81	47.34	76.62	83.61	87.66
✓			1.25	61.89	85.06	91.01	92.81	1.62	50.10	79.30	85.12	87.79
✓	✓		1.21	59.84	85.78	91.21	93.47	1.50	52.28	79.62	86.22	89.46
✓		✓	1.15	63.42	85.55	91.08	93.87	1.53	53.35	80.04	86.00	89.35
✓	✓	✓	**1.09**	**64.15**	**86.55**	**91.18**	**93.95**	**1.43**	**55.25**	**80.64**	**86.53**	**89.58**

3.3 Analytical Ablation Studies

To validate the effects of our network components, we conducted ablation experiments by augmenting the base network, with the results shown in Table 2. We observed that adding only the CAM and simply concatenating contours with baseline features significantly improved performance on the source domain (MRE decreased from 1.44 to 1.25). Further incorporating the JAM module led to notable improvements in SDR at the 2.5 mm and 3 mm thresholds on the target domain, increasing by 1.10% and 1.67%, respectively. We attribute these improvements to the interaction of features from different sources (contours and contextual), which enhanced deep anatomical features and reduced large deviations. When only the CAM and DRM were added, MRE and SDR within 1 mm improved significantly, benefiting from the direction-based regression loss optimization. Our method achieved the best performance across all module configurations.

4 Conclusion

In this paper, we propose a contour-guided cross-domain learning (CJL) framework for cephalometric landmark detection. By leveraging cranial soft/hard tissue contours as domain-invariant anatomical priors and incorporating a joint attention module, our method effectively bridges domain gaps and enhances cross-domain generalization. Additionally, by introducing an anisotropic regression module, we further improve landmark detection accuracy across cross-domain datasets. Experimental results on the multi-domain CEPHA29 and ISBI 2015 datasets demonstrate the effectiveness of our approach in maintaining anatomical plausibility and achieving state-of-the-art performance under diverse imaging conditions.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant (No.62172257), the Natural Science Foundation of Shandong Province (Major Basic Research) project under Grant (ZR2024ZD12).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arık, S.Ö., Ibragimov, B., Xing, L.: Fully automated quantitative cephalometry using convolutional neural networks. Journal of Medical Imaging **4**(1), 014501–014501 (2017)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Chen, R., Ma, Y., Chen, N., Lee, D., Wang, W.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 873–881. Springer (2019)
4. Delaire, J.: L'analyse architecturale et structurale cranio-faciale, vol. 82. Revue de Stomatologie et de Chirurgie Maxillo-faciale (1981)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Gilmour, L., Ray, N.: Locating cephalometric x-ray landmarks with foveated pyramid attention. In: Medical Imaging With Deep Learning. pp. 262–276. PMLR (2020)
7. Jacobson, A., Jacobson, R.L. (eds.): Radiographic Cephalometry: From Basics to 3-D Imaging. Quintessence Publishing, Chicago, 2 edn. (2006)

8. Jin, H., Che, H., Chen, H.: Unsupervised domain adaptation for anatomical landmark detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 695–705. Springer (2023)
9. Khalid, M.A., Zulfiqar, K., Bashir, U., Shaheen, A., Iqbal, R., Rizwan, Z., Rizwan, G., Fraz, M.M.: Cepha29: automatic cephalometric landmark detection challenge 2023. arXiv preprint arXiv:2212.04808 (2022)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
11. Lee, M., Chung, M., Shin, Y.G.: Cephalometric landmark detection via global and local encoders and patch-wise attentions. Neurocomputing **470**, 182–189 (2022)
12. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based cnns for landmark localization. Medical image analysis **54**, 207–219 (2019)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
15. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical image analysis **31**, 63–76 (2016)
16. Wu, H., Wang, C., Mei, L., Yang, T., Zhu, M., Shen, D., Cui, Z.: Cephalometric landmark detection across ages with prototypical network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 155–165. Springer (2024)
17. Wu, Q., Yeo, S.Y., Chen, Y., Liu, J.: Revisiting cephalometric landmark detection from the view of human pose estimation with lightweight super-resolution head. arXiv preprint arXiv:2309.17143 (2023)
18. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2129–2138 (2018)
19. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: You only learn once: Universal anatomical landmark detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 85–95. Springer (2021)
20. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)