DentMamba: An Anatomy-Aware Global-Local Hybrid Network for Large-Scale Multi-Class Dental Disease Detection

Kang Zhang[1,2], Shaojie Zhuang[1], Guangshun Wei[1(✉)], and Yuanfeng Zhou[1(✉)]

[1] School of Software, Shandong University, Jinan, Shandong 250100, China, yfzhou@sdu.edu.cn

[2] School of Information Engineering, Changji University, Changji, Xinjiang 831100, China, 19036@cjc.edu.com.cn

Abstract. Dental disease detection on panoramic radiographs is challenged by extreme scale variation, tiny lesions with ambiguous boundaries, and the vertical anisotropy of tooth anatomy. To improve sensitivity to small lesions while retaining computational efficiency, we propose DentMamba, a two-stage framework: a global tooth localization stage to isolate tooth ROIs and suppress background structures, and a fine-grained detector built on anatomy-aware global–local modeling. In the second stage, we introduce the AnatoMamba block with anisotropic convolutions to encode vertical root priors while avoiding the quadratic cost of global attention, and a bidirectional MultiScale Adaptive Spatial Attention Gate fusion module (MASAG) to reduce semantic misalignment during feature fusion. Furthermore, we propose a Dynamic Scale-Aware Hybrid Loss (DSAHL) that adapts the regression objective across lesion scales and aligns classification confidence with localization quality. Experiments on a large clinical panoramic dataset (20,000 images; 18,534 used for cross-validation) demonstrate that DentMamba achieves 84.6 AP with 17M parameters, with notably improved small-lesion performance (AP_S 52.1), outperforming recent detectors under a strong accuracy–efficiency trade-off. The evaluation code can be accessed at: https://github.com/zhkangzhkang/Dent_Mamba.

Keywords: Panoramic X-ray · Deep Learning · Anatomy-Aware Learning · Computer-Aided Diagnosis.

1 Introduction

Oral diseases impact approximately 3.5 billion people globally. With the growing number of patients undergoing dental treatment, X-ray panoramic radiography has been widely used in clinical practice due to its low radiation dose, broad anatomical coverage, and cost effectiveness. It plays an essential role in the screening and diagnosis of oral and maxillofacial diseases. Since manual interpretation is labor-intensive, deep learning has emerged as a critical tool for automated analysis. Early researches based on U-Net or VGG architectures

have achieved success in restoration detection [13] and implant classification [14]. YOLO-based models [1,17] have been proven effective for object detection. However, lesion detection on panoramic radiography remains challenging, as some lesions correspond to micro-scale or inconspicuous regions, or are strongly affected by class imbalance in the training data, making these regions difficult for automated detection algorithms to identify.

To address these challenges, researchers have optimized network structures. Strategies include enhancing downsampling in DC-YOLO [10], incorporating Transformer mechanisms for portable devices [7], and employing dual-stage networks to improve recall rates [8]. Recently, complex architectures have been proposed for multi-scale variations and diverse lesion morphology. Specifically, DVCTNett [11] integrates dental details via Gated Cross-View Attention, while MOQAMt [2] enhances robustness by addressing long-tail distributions.

Despite these improvements, the challenges of class imbalance and small lesion detection persist, severely limiting the reliability of automated detection algorithms in clinical applications. CNN-based methods are limited by their fixed-size receptive fields, making it difficult to simultaneously adapt to lesions of varying sizes. Transformer-based approaches, while capable of capturing global contextual information, typically incur substantial computational costs, which limits their practicality for clinical applications. Furthermore, existing models often neglect specific anatomical priors, such as the vertical anisotropy of teeth, leading to detection errors. State Space Models (SSMs) like Mamba [5] offer linear complexity with strong global modeling capabilities, presenting a promising solution.

To address the challenges above, we propose DentMamba, a novel anatomy-aware detection framework. By integrating Mamba into multi-scale feature learning, DentMamba effectively balances global semantic understanding with local detail extraction. Our main contributions are as follows.

1. **Large-Scale Dataset Creation:** We have constructed a large-scale panoramic dataset with over 20,000 annotated images across 12 lesion categories, providing a valuable benchmark for future dental research.
2. **Anatomical Feature Fusion:** To address vertical anisotropy and feature misalignment, we introduce the AnatoMamba Block and MASAG module, enhancing the detection of concealed lesions through anatomical feature integration and bidirectional gating.
3. **Scale-Invariant Detection:** We propose the Dynamic Scale-Aware Hybrid Loss (DSAHL), which combines Scale-Aware Regression and NWD-based focal loss, ensuring precise detection across a wide range of lesion scales.

2 Method

2.1 Overview of the Proposed Framework

As shown in Fig. 1, DentMamba is a hierarchical framework mirroring the clinical coarse-to-fine diagnostic workflow. The feature encoder follows the efficient

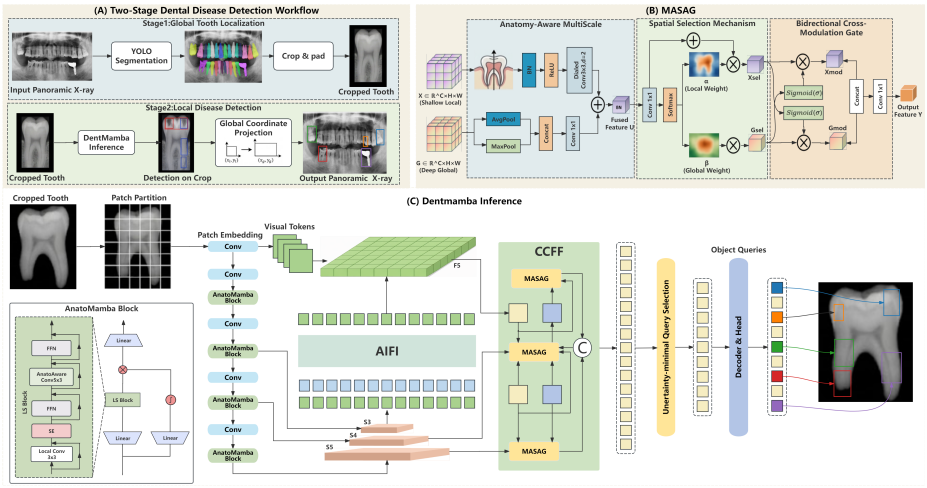

Fig. 1. The overall framework of the proposed DentMamba.

hybrid design of RT-DETR [20], where AIFI and CCFF refer to Attention-based Intra-scale Feature Interaction and CNN-based Cross-scale Feature Fusion, respectively. This design addresses the unique challenges of dental panoramic radiographs: extreme morphological and scale variations, highly imbalanced disease distributions, and the coexistence of prominent global dentition alongside minute, ill-defined lesions (e.g., caries, periodontitis).

To circumvent information loss from downsampling full-resolution images, we decompose the task into two stages.

Stage 1: Global Tooth Localization. As illustrated in Fig. 1, the first stage identifies tooth-level regions from the panoramic radiograph and generates anatomically meaningful ROIs for subsequent fine-grained analysis. Unlike tooth numbering, this stage only requires each tooth region to be sufficiently covered by the cropped ROI, thereby reducing background interference and tooth-scale variation before lesion detection. On 600 panoramic radiographs containing 17,591 tooth instances, the module detected 17,564 teeth at IoU=0.5, achieving a tooth-level detection rate of 99.8%, suggesting that the first-stage error is negligible for the subsequent detection stage.

Stage 2: Local Disease Detection. The cropped tooth images are processed by DentMamba, a novel visual token-based architecture tailored for dental anatomy (Fig. 1C). The input is partitioned into visual tokens, processed through a hierarchical backbone of stacked AnatoMamba Blocks. These blocks utilize local-spatial designs and anatomy-aware convolutions to efficiently capture vertical root dependencies, avoiding the quadratic complexity of global attention.

2.2 AnatoMamba Block

Dental structures like roots, canals, and periodontal ligaments exhibit vertical anisotropy. Standard isotropic convolutions (e.g., 3×3) are ineffective for capturing these elongated features, and global attention mechanisms incur high computational cost. To address this, we propose the AnatoMamba Block (see Fig. 1C), which integrates anatomical priors into a gated architecture for linear complexity and better contextual modeling.

Our module balances structural integrity and local detail through feature extraction. It first uses a 3×3 depth-wise convolution to capture high-frequency textures like enamel caries. Then, an Anatomy-Aware Convolution with a 5×3 kernel aligns with the vertical structure of teeth, expanding the receptive field to capture root structures while minimizing interference from adjacent teeth.

To eliminate irrelevant anatomical noise without the computational overhead of self-attention, we utilize a Gated Linear Unit (GLU) framework based on State Space Models (SSMs). The feature flow is modulated by a gating branch that prioritizes lesion-related signals. Given the refined anatomical features H_{anat} and input sequence X, operation is applied.

2.3 MultiScale Adaptive Spatial Attention Gate (MASAG)

In hierarchical detection, traditional fusion methods (e.g., summation or concatenation) suffer from semantic misalignment. Deep features capture global semantics but lack precise localization, while shallow features preserve boundaries but lack context. To address this, we propose the MASAG, shown in Fig. 1B, which accounts for vertical anisotropy in dental structures.

Unlike conventional spatial attention, MASAG creates a guidance map by fusing anatomy-aware local details with global context. Here, X and G denote the aligned local and global features with the same size $\mathbb{R}^{B\times C\times H\times W}$. We first extract vertically-aligned local features X_{local} from X using a 5×3 anisotropic convolution, followed by dilated convolution. Simultaneously, global descriptors G_{global} are obtained from G through spatial pooling. The joint guidance map is:

$$U = \mathrm{BN}(X_{local} + G_{global}), \tag{1}$$

where BN refers to Batch Normalization.

A bidirectional gating mechanism uses the guidance map U to enable cross-hierarchy message passing. Specifically, the spatial gates α and β are generated from U by softmax and broadcast along the channel dimension during element-wise modulation. The modulated features X_{mod} and G_{mod} are:

$$\begin{aligned} X_{mod} &= (X \odot \alpha) \odot \sigma(G \odot \beta), \\ G_{mod} &= (G \odot \beta) \odot \sigma(X \odot \alpha), \end{aligned} \tag{2}$$

where $\odot$ denotes element-wise multiplication. Finally, the modulated features are concatenated and fused through a 1×1 convolution, resulting in the output:

$$Y = \mathrm{Conv}_{1\times1}([X_{mod}; G_{mod}]). \tag{3}$$

This bidirectional modulation reduces semantic ambiguity and improves detection of scale-variant dental anomalies.

2.4 Dynamic Scale-Aware Hybrid Loss (DSAHL)

Dental lesions exhibit significant scale variations, from small caries to large restorations. Traditional IoU-based metrics struggle with minor positional deviations in small targets, leading to unstable optimization.

To address this issue, we introduce the DSAHL, which combines Gaussian distribution modeling with adaptive optimization. The bounding boxes are modeled as 2D Gaussians, $\mathcal{N}(\mu, \Sigma)$. By incorporating the Normalized Wasserstein Distance (NWD), we shift the overlap measure from a "rigid pixel-based intersection" to a "distribution-based similarity," providing continuous and smooth gradient guidance, particularly for small lesions:

$$\mathrm{NWD}(\mathcal{N}_p, \mathcal{N}_g) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_p, \mathcal{N}_g)}}{C}\right) \tag{4}$$

Here, each box $B = (c_x, c_y, w, h)$ is modeled as $\mathcal{N}((c_x, c_y)^\top, \mathrm{diag}(w^2/12, h^2/12))$, and W_2^2 denotes the squared Wasserstein distance between predicted and ground-truth boxes.

To focus on object size, we introduce a scale-adaptive weight $\alpha_i = 1 - \exp(-\sqrt{w_i h_i}/\tau)$, where small targets get a weight close to 0 and large ones close to 1, smoothly transitioning to the boundary-sensitive GIoU loss.

$$\mathcal{L}_{reg} = \frac{1}{N_{pos}} \sum_i \left[(1-\alpha_i)(1-\mathrm{NWD}_i) + \alpha_i \mathcal{L}_{\mathrm{GIoU}}(P_i, G_i)\right]. \tag{5}$$

Additionally, to address the misalignment between classification confidence and localization, we propose the NWD-Aware Quality Focal Loss (NQFL), which replaces hard labels with soft labels using the NWD metric to guide the classification probability.

$$\mathcal{L}_{cls} = -\sum_i |q_i - p_i|^\beta \left(q_i \log(p_i) + (1-q_i)\log(1-p_i)\right). \tag{6}$$

The total optimization objective is the weighted sum of these components:

$$\mathcal{L}_{total} = \lambda_{reg}\mathcal{L}_{reg} + \lambda_{cls}\mathcal{L}_{cls}. \tag{7}$$

3 Experiments

3.1 Dataset and Evaluation Metrics

We collected 20,000 high-resolution (2610×1560) dental panoramic radiographs from a clinical partner. Data quality was ensured by a strict hierarchy of 8 annotators, 3 auditors, and 1 expert, featuring multi-round reviews and a 90%

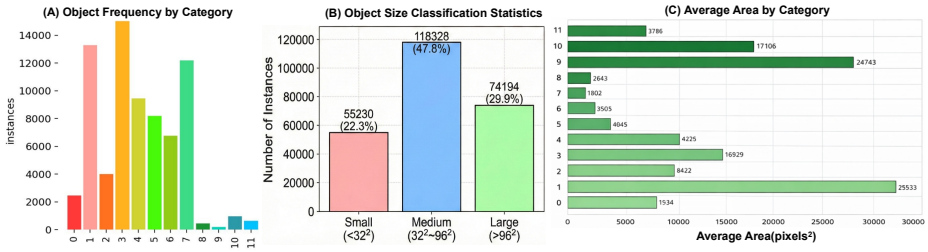

Fig. 2. Dataset Comprehensive Analysis Dashboard.

quality threshold. For experiments, 18,534 validated images were utilized for model training and evaluation via K-fold cross-validation.

The annotation protocol covers 12 clinically relevant dental disease categories in panoramic radiographs. Specifically, the annotated categories include impacted tooth, implant, fixed bridge, crown restoration, filling, root canal filling, caries, dental calculus, post-and-core restoration, residual crown, residual root, and periapical periodontitis.

As shown in Fig. 2, the dataset exhibits a clear long-tailed distribution and substantial target-scale variation. Common categories such as filling appear much more frequently than rare categories such as periapical periodontitis, while small targets such as caries, dental calculus, and periapical radiolucency are often difficult to detect due to their limited size and ambiguous boundaries. These characteristics motivate our anatomy-aware backbone and scale-adaptive loss function for robust dental disease detection.

3.2 Implementation Details

Our method is implemented in PyTorch on four NVIDIA RTX 3090 GPUs. The model is trained from scratch for 100 epochs (batch size 32) using the AdamW optimizer (initial $lr = 1e^{-4}$, weight decay 0.05, cosine schedule). Input images are resized to 416×416. Data augmentation includes Mosaic and Mixup, while vertical flipping is explicitly disabled to maintain the dental root's vertical anisotropy. For the loss function, we set $\lambda_{reg} = 5.0$, $\lambda_{cls} = 1.0$, and the NWD constant $C = 12.8$.

3.3 Comparison with State-of-the-Art Methods

To evaluate DentMamba's performance, we compared it with CNN and Transformer-based detectors using two large-scale datasets: our private clinical dataset and the public OralXrays-9 benchmark. The detailed results are summarized in **Table 1**.

Efficiency vs. Transformers. While Transformers like Co-DINO and MO-QAM offer high performance, they require 66M parameters. In contrast, DentMamba achieves state-of-the-art accuracy with only **17M** parameters, a 3.8× reduction, offering a much better accuracy-efficiency trade-off for clinical use.

Table 1. Comprehensive comparison with state-of-the-art methods on our Clinical Dataset and the public OralXrays-9 Benchmark. The best results are highlighted in **bold**, and the second-best are underlined.

Method	Params	Our Clinical Dataset						OralXrays-9 Benchmark					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CNN-based Methods													
Faster R-CNN [4]	41M	78.5	85.2	79.1	32.5	55.4	80.2	69.1	91.4	72.8	19.1	62.3	69.3
ATSS [19]	32M	79.2	86.1	80.3	34.1	56.8	81.5	69.6	91.2	73.0	17.6	63.5	70.4
YOLOX [3]	99M	81.0	88.5	82.4	38.5	59.2	83.1	68.9	90.1	72.1	19.4	63.0	69.0
Sparse R-CNN [15]	106M	80.5	87.2	81.0	36.8	58.5	82.0	69.0	90.2	71.9	19.6	61.1	69.7
CFinet [16]	49M	81.8	89.0	82.8	40.2	60.1	83.5	69.8	90.5	72.5	20.1	62.3	70.5
Transformer-based Methods													
DETR [20]	42M	77.8	84.5	78.0	30.2	54.1	79.5	62.3	90.2	68.9	12.1	55.7	64.0
Deformable-DETR [21]	40M	82.1	89.4	83.1	41.5	61.2	84.0	68.7	91.2	73.0	17.6	61.2	70.1
Conditional-DETR [12]	46M	81.5	88.8	82.5	40.8	60.5	83.2	65.0	89.9	70.0	15.8	57.6	66.1
DAB-DETR [9]	55M	82.8	90.1	84.2	42.6	62.0	84.8	67.4	91.3	72.3	17.8	60.3	68.6
DINO [18]	56M	83.5	91.5	85.8	44.7	63.2	86.2	69.9	89.7	71.9	18.9	63.3	70.9
Co-DINO [22]	66M	83.2	91.8	86.1	44.8	63.8	86.1	70.9	90.2	72.4	19.7	64.8	71.3
Salience-DETR [6]	56M	83.9	92.2	86.8	46.2	64.5	86.0	71.0	91.1	73.5	19.5	64.9	71.5
MOQAM [2]	66M	84.3	92.5	87.1	45.8	64.0	86.3	73.5	92.7	74.9	21.2	68.1	74.0
DentMamba (Ours)	**17M**	**84.6**	**95.1**	**93.2**	**52.1**	**65.8**	**88.4**	**74.6**	**93.1**	**75.2**	**24.5**	**69.4**	**74.9**

Superiority and Generalization. DentMamba outperforms standard CNNs (e.g., Faster R-CNN) by a large margin. On the OralXrays-9 benchmark, it sets a new state-of-the-art of **74.6%** AP, surpassing MOQAM by **+1.1%**, showing strong generalization across different datasets and imaging conditions.

Efficacy in Small Lesion Detection. Leveraging the AnatoMamba Block and DSAHL loss, DentMamba excels at detecting small and ambiguous lesions. It improved AP_S by **5.9%** on our clinical dataset (reaching **52.1%**) and by **3.3%** (reaching **24.5%**) on OralXrays-9, demonstrating that integrating anatomical priors is more effective than generic global attention.

3.4 Qualitative Analysis

To provide a comprehensive evaluation of DentMamba, we integrate detection results and interpretability visualizations in Fig. 3. The figure displays the original input (a), ground truth (b), our detection results (c), and a comparison of attention heatmaps (d)-(e).

Detection Precision and Structural Alignment. DentMamba (c) shows high consistency with the Ground Truth (b), particularly in challenging cases like minute lesions (e.g., early caries). The model effectively suppresses background noise (a) and tightly aligns with vertical dental roots, addressing vertical anisotropy.

Interpretability via Grad-CAM. The baseline heatmap (d) shows scattered attention, influenced by irrelevant features, while DentMamba's heatmap (e) focuses sharply on dental structures and lesion boundaries. This highlights the effectiveness of our 'See Large, Focus Small' mechanism in prioritizing anatomical features.

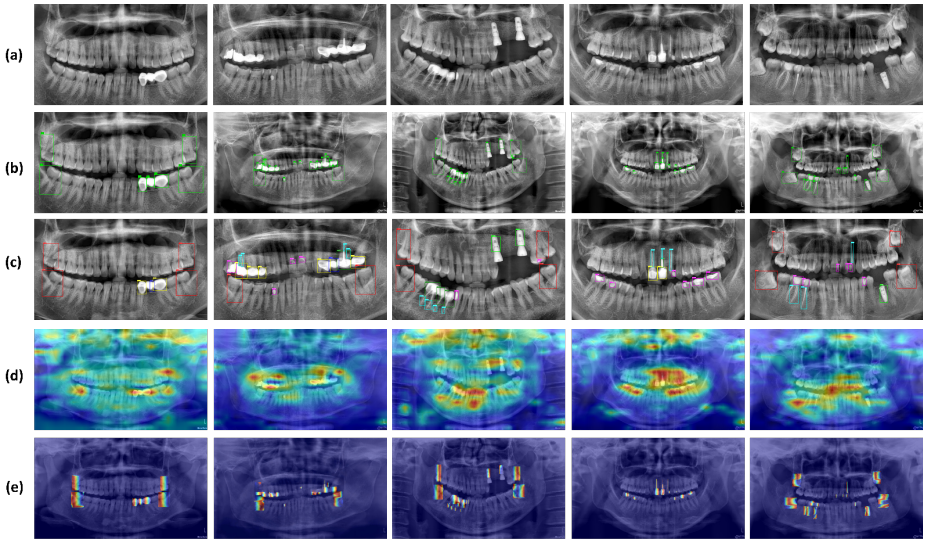

Fig. 3. Qualitative detection results and Grad-CAM visualization.

Table 2. Component-wise ablation study. GA-LS: Gated Anatomy-Aware LS MambaOut Backbone; MASAG: MultiScale Adaptive Spatial Attention Gate;DSAHL: Dynamic Scale-Aware Hybrid Loss.

Ablation Aspect	Variant / Configuration	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Components	Baseline	81.4	91.2	88.4	45.1	62.3	85.4
	+ AnatoMamba	82.9	92.5	90.1	47.3	63.6	86.1
	+ AnatoMamba & MASAG	84.1	93.8	91.5	49.8	65.4	87.2
	+ All (Ours Full)	**84.6**	**95.1**	**93.2**	**52.1**	**65.8**	**88.4**
Kernel Shape	3×3 (Standard Isotropic)	84.1	94.5	92.7	49.5	64.0	86.3
	5×5 (Large Isotropic)	84.4	94.8	92.9	50.2	65.1	87.0
	3×5 (Horizontal Mismatch)	83.2	93.5	91.6	48.9	61.8	85.0
	5×3 (Vertical / Default)	**84.6**	**95.1**	**93.2**	**52.1**	**65.8**	**88.4**
Fusion Strategy	Element-wise Summation	83.7	94.1	92.2	49.2	62.7	85.3
	Concatenation	84.2	94.6	92.8	49.6	64.2	86.6
	Unidirectional ($G \rightarrow X$)	84.4	94.9	93.0	50.5	65.3	87.0
	Bidirectional (Default)	**84.6**	**95.1**	**93.2**	**52.1**	**65.8**	**88.4**

3.5 Ablation Studies

We validate the effectiveness of DentMamba and its core designs through comprehensive ablation experiments on the dataset, as detailed in Table 2.

Component Contribution. As shown in the top section of Table 2, the integration of the AnatoMamba Block, MASAG, and DSAHL consistently improves detection performance. Notably, incorporating DSAHL boosts AP_S by +2.3%, highlighting its role in enhancing sensitivity to small, hard-to-detect lesions.

Design Validation. Internal analysis supports our architectural choices: (1) **Kernel Shape:** The 5×3 anisotropic kernel outperforms standard isotropic kernels and horizontal mismatch kernels, effectively capturing vertical anatom-

ical features. (2) **Fusion Strategy:** Bidirectional Cross-Modulation is superior to simpler fusion methods, improving lesion localization by balancing global semantics with local details.

4 Conclusion

We proposed DentMamba, a bio-inspired framework addressing the anatomical complexity of dental radiography. By synergizing the vertical-aware AnatoMamba backbone, bidirectional MASAG fusion, and scale-adaptive DSAHL, our method effectively balances global structural perception with minute lesion detection. Extensive experiments on 20,000 images confirm its superior accuracy-efficiency trade-off against state-of-the-art methods. Future work will extend these biological priors to 3D CBCT analysis and real-time edge deployment.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62572284 and 62502273), the Natural Science Foundation of Shandong Province (Major Basic Research Project, Grant No. ZR2024ZD12; Youth Fund, Grant No. ZR2025QC1558), the High-Level Innovation Teams of Universities and Research Institutes of Jinan (Grant No. 202534004), and the Natural Science Foundation of Xinjiang Uygur Autonomous Region (Grant No. 2025D01C84).

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

References

1. Bonfanti-Gris, M., Herrera, A., Paraíso-Medina, S., Alonso-Calvo, R., Martínez-Rus, F., Pradíes, G.: Performance evaluation of three versions of a convolutional neural network for object detection and segmentation using a multiclass and reduced panoramic radiograph dataset. Journal of dentistry **144**, 104891 (2024)
2. Chen, B., Fu, S., Fang, X., Cai, J., Zhang, B., Lu, M., Liu, Y.: Oralxrays-9: Towards hospital-scale panoramic x-ray anomaly detection via personalized multi-object query-aware mining. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15570–15579 (2025)
3. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
6. Hou, X., Liu, M., Zhang, S., Wei, P., Chen, B.: Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17574–17583 (2024)
7. Jiang, H., Zhang, P., Che, C., Jin, B.: Rdfnet: A fast caries detection method incorporating transformer mechanism. Computational and Mathematical Methods in Medicine **2021**(1), 9773917 (2021)

8. Jiao, X., Gao, S., Huang, F., Dou, W., Zhou, Y., Zhang, C.: An improved algorithm for full-mouth lesion detection based on yolov8. Graphical Models **142**, 101302 (2025)
9. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
10. Lou, H., Duan, X., Guo, J., Liu, H., Gu, J., Bi, L., Chen, H.: Dc-yolov8: small-size object detection algorithm based on camera sensor. Electronics **12**(10), 2323 (2023)
11. Luo, T., Wu, H., Yang, T., Shen, D., Cui, Z.: Adapting foundation model for dental caries detection with dual-view co-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 46–55. Springer (2025)
12. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3651–3660 (2021)
13. Oztekin, F., Katar, O., Sadak, F., Aydogan, M., Yildirim, T.T., Plawiak, P., Yildirim, O., Talo, M., Karabatak, M.: Automatic semantic segmentation for dental restorations in panoramic radiography images using u-net model. International Journal of Imaging Systems and Technology **32**(6), 1990–2001 (2022)
14. Park, J.H., Moon, H.S., Jung, H.I., Hwang, J., Choi, Y.H., Kim, J.E.: Deep learning and clustering approaches for dental implant size classification based on periapical radiographs. Scientific reports **13**(1), 16856 (2023)
15. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14454–14463 (2021)
16. Yuan, X., Cheng, G., Yan, K., Zeng, Q., Han, J.: Small object detection via coarse-to-fine proposal generation and imitation learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6317–6327 (2023)
17. Yüksel, A.E., Gültekin, S., Simsar, E., Özdemir, Ş.D., Gündoğar, M., Tokgöz, S.B., Hamamcı, İ.E.: Dental enumeration and multiple treatment detection on panoramic x-rays using deep learning. Scientific reports **11**(1), 12342 (2021)
18. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations
19. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9759–9768 (2020)
20. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2024)
21. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations
22. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6748–6758 (2023)