Adapting Foundation Model for Dental Caries Detection with Dual-View Co-Training

Tao Luo[1*], Han Wu[1,2*], Tong Yang[5], Dinggang Shen[1,3,4], and Zhiming Cui[1(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
cuizhm@shanghaitech.edu.cn
[2] Lingang Laboratory, Shanghai, China
[3] Shanghai United Imaging Intelligence Co. Ltd., Shanghai, China
[4] Shanghai Clinical Research and Trial Center, Shanghai, China
[5] Shanghai Linkedcare Information Technology Co., Ltd., Shanghai, China

Abstract. Accurate dental caries detection from panoramic X-rays plays a pivotal role in preventing lesion progression. However, current detection methods often yield suboptimal accuracy due to subtle contrast variations and diverse lesion morphology of dental caries. In this work, inspired by the clinical workflow where dentists systematically combine whole-image screening with detailed tooth-level inspection, we present **DVCTNet**, a novel **D**ual-**V**iew **C**o-**T**raining network for accurate dental caries detection. Our DVCTNet starts with employing automated tooth detection to establish two complementary views: a global view from panoramic X-ray images and a local view from cropped tooth images. We then pretrain two vision foundation models separately on the two views. The global-view foundation model serves as the detection backbone, generating region proposals and global features, while the local-view model extracts detailed features from corresponding cropped tooth patches matched by the region proposals. To effectively integrate information from both views, we introduce a Gated Cross-View Attention (GCV-Atten) module that dynamically fuses dual-view features, enhancing the detection pipeline by integrating the fused features back into the detection model for final caries detection. To rigorously evaluate our DVCTNet, we test it on a public dataset and further validate its performance on a newly curated, high-precision dental caries detection dataset, annotated using both intra-oral images and panoramic X-rays for double verification. Experimental results demonstrate DVCTNet's superior performance against existing state-of-the-art (SOTA) methods on both datasets, indicating the clinical applicability of our method. Our code and labeled dataset are available at https://github.com/ShanghaiTech-IMPACT/DVCTNet.

Keywords: Caries Detection · Foundation Model · Dual-View Co-Training · Gated Cross-View Attention.

[*] indicates equal contribution.

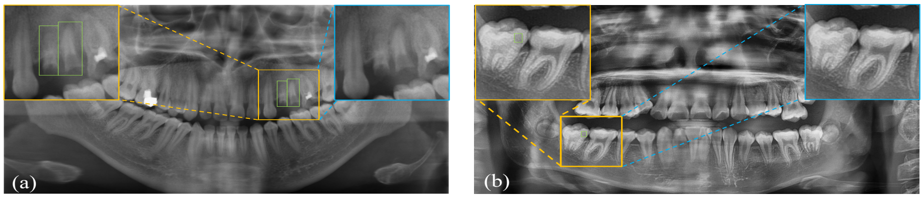

Fig. 1: Comparison of annotated cases between recent public dataset [1] (a), and our dataset (b). Dashed lines link zoomed views: blue section shows the original image, orange section highlights the same area with overlaid annotated boxes.

1 Introduction

Dental caries is one of the most common oral diseases [8,17], which may lead to irreversible teeth damage. Accurate caries detection from panoramic X-rays is crucial to prevent lesion progression [11,5]. However, manually detecting caries from panoramic X-rays remains time-consuming and labor-intensive due to their radiographic diversity and inconspicuous signs. With the development of deep learning [3,18,22], existing methods have shown great success in automated caries detection from panoramic X-rays [7,20,12]. For example, Zhu et al. [23] incorporated a full-scale axial attention module, achieving notable performance in segmenting caries. Wang et al. [16] proposed a lightweight region of interest pruning method that effectively improves caries detection with a novel label assignment head. Chen et al. [1](referred to as FPCL) proposed a proposal-prototype contrastive learning method based on a bidirectional FPN for caries detection.

Despite encouraging performance, current approaches still face two major limitations. 1) Learning-based methods rely heavily on the existence of well-annotated datasets, while current public datasets [1,21] fail to provide a high-quality golden-standard dataset where clinical practice addresses uncertain dental caries through dual-modality verification with panoramic X-rays and intraoral images [14]. Additionally, they tend to over-annotate caries areas as shown in Fig. 1. 2) Existing methods directly adapt generic models from computer vision, which are either limited to single-view analysis at the image level or fail to align with the comprehensive clinical diagnostic workflow, leading to suboptimal results [9,16].

In this paper, we introduce DVCTNet, a novel dual-view co-training framework that leverages foundation models for dental caries detection. Specifically, DVCTNet generates dual-view images: panoramic X-rays and cropped tooth images, to align with the clinical workflow, where dentists typically combine whole-image screening with detailed tooth-level inspection for accurate diagnosis. Our DVCTNet pretrains two foundation models on global and local views, yielding dual-view image encoders on the collected large-scale, unlabeled dataset. The global view encoder serves as the backbone for the object detector, generating region proposals and global features, while the local view encoder extracts fine-

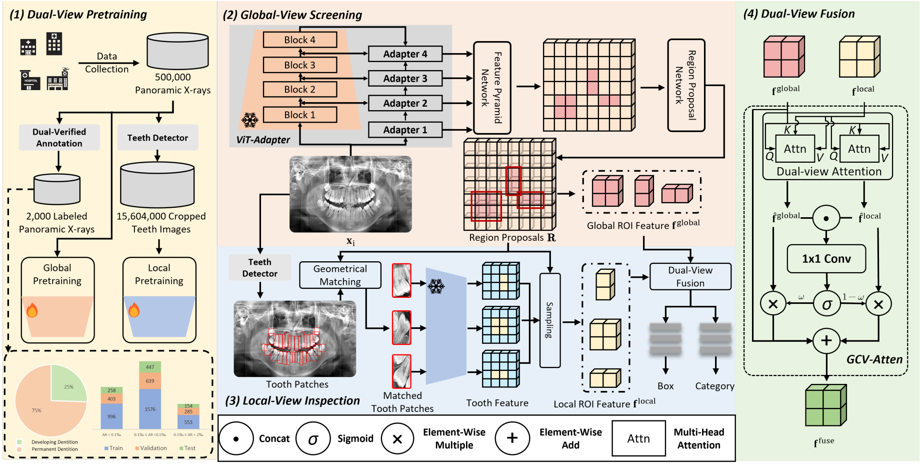

Fig. 2: An overview of the proposed DVCTNet for dental caries detection.

grained features from cropped tooth patches matched with the region proposals, complementing the global representation. To effectively integrate dual-view features, a Gated Cross-View Attention (GCN-Atten) mechanism is developed to dynamically fuse the features, enhancing the detection pipeline by reincorporating the fused features into the model for final caries detection. We conducted extensive experiments on two datasets: 1) a publicly available dataset from [1], and 2) a high-quality and newly-collected dataset dual-verified with intra-oral images and panoramic X-rays, and cross-validated by four dental experts. Experimental results demonstrate that DVCTNet outperforms existing state-of-the-art (SOTA) methods on both datasets, indicating strong clinical applicability. In summary, **our contributions can be summerized as followed**: 1) A novel dual-view co-training mechanism that leverages foundation models to extract and utilize complementary features from global panoramic and local tooth-specific views for accurate dental caries detection. 2) A Gated Cross-View Attention mechanism that dynamically fuses features from dual views, optimizing the integration of contextual and detail-oriented information. 3) We present the first high-precision benchmark dataset for dental caries detection with annotations double-verified through intra-oral images and panoramic X-rays.

2 Method

The overview pipeline of DVCTNet is shown in Fig. 2. First, we establish complementary views and pretrain foundation models on both panoramic X-rays and tooth images (Sec. 2.1). Next, we integrate these pretrained models through dual-view co-training for caries detection, combining global screening with local inspection (Sec. 2.2). Finally, we introduce a Gated Cross-View Attention mechanism for dual-view fusion (Sec. 2.3).

2.1 Dual-View Pretraining

Our method begins with the establishment of dual views that closely follow the clinical diagnostic workflow for caries detection. Given an input panoramic X-ray image $\mathbf{X} \in \mathbb{R}^{H \times W}$, we employ a well-developed tooth detector $\mathcal{D}$ (i.e., tooth detector form [11]) to generate bounding boxes around individual teeth and crop each tooth, resulting in local tooth images $\mathbf{T} = \{\mathbf{t}_i \in \mathbb{R}^{h \times w}\}_{i=1}^{N}$, where N is the number of teeth present in the X-ray. $H \times W$ and $h \times w$ are the resolutions of the X-ray and cropped tooth image, respectively. This process yields dual views with complementary information: a global view of the entire panoramic X-ray image $\mathbf{X}$ and a corresponding local view $\mathbf{T}$ consisting of individual tooth images.

As shown in Fig.2(1), to effectively utilize the large amount of unlabeled data available in dental imaging and to better capture the intrinsic features of both global and local views, we introduce a dual-view pretraining stage based on DINOv2 [13], a state-of-the-art self-supervised learning framework. DINOv2 employs a teacher-student distillation mechanism with momentum updating, allowing the model to learn consistent representations across different augmentations of the same image. In our study, for the **global view pretraining**, we configure the ViT-B model with a patch size of 14×14 and input resolution of 518×518 pixels. During training, the model uses multi-scale argumentation with different crops, e.g., 518×518 and 224×224 pixels, enabling the model to capture hierarchical information in the panoramic X-rays. For the **local view pretraining**, we use the ViT-B model but with smaller crop dimensions appropriate for tooth-level analysis, e.g., 112×112 and 98×98 pixels. This configuration is specifically designed to effectively capture the fine-grained dental information presented in individual tooth images.

2.2 Dual-View Co-Training

Following the dual-view pretraining stage, we integrate the pretrained models into dental caries detection. The core idea of the dual-view co-training is to simulate the real-world clinical workflow, where dentists combine global-view screening from the entire panoramic X-ray images with local-view inspection at the tooth level for an accurate diagnosis.

Global-View Screening As shown in Fig. 2(2), we develop a multi-scale feature extraction framework based on ViTAdapter [2]. The ViTAdapter extends the original Vision Transformer architecture by introducing deformable attention modules between transformer blocks, enabling the model to capture multi-scale features more effectively. Specifically, we apply these adapters among the original ViT blocks of the image-level encoder pretrained in Sec.2.1. This design allows our model to generate multi-scale feature maps with dimensions consistent with traditional detection backbones [6]. These feature maps containing hierarchical semantic information are then fused through a Feature Pyramid Network (FPN) to combine multi-scale information. A Region Proposal Network (RPN) then generates region proposals $\mathbf{R} = \{\mathbf{r}_j, \mathbf{f}_j^{\text{global}}\}_{j=1}^{M}$ from these fused features, where each $\mathbf{r}_j \in \mathbb{R}^4$ represents the bounding box of the proposal (center coordinates, width, and height), and $\mathbf{f}_j^{\text{global}} \in \mathbb{R}^{7\times 7}$ indicates the corresponding ROI features form the global view using RoIAlign.

Local-View Inspection As shown in Fig. 2(3), to ensure our model simultaneously capture the fine-grained details from individual teeth, we perform geometrical matching between each region proposal bounding box $\mathbf{r}_j$ and the most appropriate tooth image $\mathbf{t}_{i^*} \in \mathbf{T}$ detected from the entire panoramic X-ray. The matching is based on Intersection over Detection (IoD) overlap, defined as:

$$i^* = \arg\max_{\mathbf{t}_i \in \mathbf{T}} \text{IoD}(\mathbf{r}_j, \mathcal{B}(\mathbf{t}_i)), \tag{1}$$

where $\mathcal{B}(\mathbf{t}_i)$ denotes the bounding box of the cropped tooth image $\mathbf{t}_i$ in the original panoramic X-ray coordinate system, and IoD is calculated as:

$$\text{IoD}(\mathbf{r}_j, \mathcal{B}(\mathbf{t}_i)) = \frac{\text{Area}(\mathbf{r}_j \cap \mathcal{B}(\mathbf{t}_i))}{\mathcal{B}(\mathbf{t}_i)}. \tag{2}$$

Once the matching tooth image $\mathbf{t}_{i^*}$ is identified, it is passed through the pre-trained local-view encoder (introduced in Sec.2.1) to extract detailed tooth-level features, and crop the corresponding ROI feature $\mathbf{f}_{i^*}^{\text{local}} \in \mathbb{R}^{7\times 7}$ from the local view.

2.3 Dual-View Fusion

To effectively integrate the complementary information from both views, we use a Gated Cross-View Attention (GCV-Atten) fusion module. This module dynamically fuses features from the global and local views based on their relevance to caries detection.

As illustrated in Fig. 2(4), given a ROI feature $\mathbf{f}_j^{\text{global}}$ from the global view and $\mathbf{f}_{i^*}^{\text{local}}$ from the local view, we first compute a dual-direction scaled dot-product attention mechanism between them. This allows each view to attend to relevant information in the other view. The attention mechanism is formulated as:

$$\hat{\mathbf{f}}_j^{\text{global}} = \text{Softmax}\left(\frac{\mathbf{f}_j^{\text{global}} \cdot {\mathbf{f}_{i^*}^{\text{local}}}^{\text{T}}}{\sqrt{d_k}}\right) \cdot \mathbf{f}_j^{\text{global}}, \tag{3}$$

$$\hat{\mathbf{f}}_{i^*}^{\text{local}} = \text{Softmax}\left(\frac{\mathbf{f}_{i^*}^{\text{local}} \cdot {\mathbf{f}_j^{\text{global}}}^{\text{T}}}{\sqrt{d_k}}\right) \cdot \mathbf{f}_{i^*}^{\text{local}}, \tag{4}$$

where $\hat{\mathbf{f}}_j^{\text{global}}$ and $\hat{\mathbf{f}}_{i^*}^{\text{local}}$ are the attention-weighted features, and $\sqrt{d_k}$ is the scaling factor. After computing the attention-weighted features, we concatenate them and pass them through a 1D convolutional layer to reduce the channel dimension. Then, a sigmoid activation is followed to compute an attention weight ω:

$$\omega = \text{Sigmoid}(\text{Conv1D}(\text{Concat}(\hat{\mathbf{f}}_j^{\text{global}}, \hat{\mathbf{f}}_{i^*}^{\text{local}}))). \tag{5}$$

In this way, the final fused ROI features can be computed as a weighted combination of the attention-weighted features from both views, with an additional residual connection from the global view features $\mathbf{f}_j^{\text{global}}$:

$$\mathbf{f}_j^{\text{fuse}} = \omega \cdot \hat{\mathbf{f}}_j^{\text{global}} + (1-\omega) \cdot \hat{\mathbf{f}}_{i^*}^{\text{local}} + \mathbf{f}_j^{\text{global}}. \tag{6}$$

The residual connection from the global view ensures the global contextual information is preserved even after fusion. And the weighted combination allows the model to adaptively balance the contributions from both views based on their relevance to the current region. Finally, the fused feature $\mathbf{f}_j^{\text{fuse}}$ is fed back to the detection head for final dental caries classification and bounding box regression.

3 Experiments

3.1 Dataset and Evaluation Metrics

AAAI Dataset FPCL [1] collected and released a CariesXrays dataset which covers 6,000 panoramic dental X-ray images, with a total of 13,783 instances of dental caries. The whole dataset is randomly divided into 4,800/1,200 for training and testing, respectively, following the original split [1] for fair comparison.

DVCT Dataset We collected a new benchmark dataset, named the DVCT dataset, which has a total number of 500,000 panoramic X-ray images, acquired from eight clinical centers, of which 498,000 remain unlabeled and 2,000 are annotated following the cross-verification by four experienced dental radiologists with over ten years of expertise, covering 5311 instances of dental caries and across dentition at different stage, as shown in Fig. 2(1). The labeled set is randomly divided into 1500/300/200 for training, validation and testing, respectively. Our new dataset has two advantages over existing ones: 1) a higher-quality golden-standard annotated with dual verification from both panoramic X-ray images and intra-oral images and also a broader coverage of subjects across different age groups and 2) a large-scale unlabeled dataset, enabling self-supervised pretaining, especially for foundation model, which can be applied for any downstream task related to dental panoramic X-ray analysis.

Evaluation Metrics For a consistent comparison, our evaluation primarily focuses on reporting the average precision (%) across all benchmark datasets following previous work [1], where we adopt the standard AP metrics under various Intersection over Union (IoU) thresholds, ranging from 0.5 to 0.95.

Table 1: Quantitative results on AAAI dataset and our DVCT dataset.

Methods	**Backbone**	**AAAI Dataset**			**DVCT Dataset**		
		AP	**AP$_{50}$**	**AP$_{75}$**	**AP**	**AP$_{50}$**	**AP$_{75}$**
RetinaNet [10]	ResNet-50	13.0	30.5	10.2	11.1	32.9	3.3
YOLOX [4]	CSPDarkNet	40.5	81.3	36.1	15.4	42.4	8.4
DINO [19]	Transformer	37.8	75.3	29.4	22.2	50.4	14.4
Faster RCNN [15]	ResNet-50	39.9	78.0	37.8	14.3	30.8	9.8
FPCL [1]	ResNet-50	48.2	84.1	50.6	17.0	42.7	10.2
DVCTNet	DINOv2-ViT-B	**48.9**	**84.7**	**52.2**	**31.3**	**57.4**	**31.9**

3.2 Implementation Details

We implemented our dual-view architecture using ViT-B (14×14 patch size) for both branches. The global view pretraining utilized 518×518 global and 224×224 local crops, while the tooth-view operated at 112×112 global and 98×98 local crops to capture fine-grained dental structures. For caries detection, we employed ViTAdapter [2] for multi-scale feature extraction at pretraining resolution, while maintaining a frozen pretrained ViT encoder with feature projection alignment in the tooth-view branch. The whole network was trained with AdamW optimizer where learning rate, weight decay, training epochs, and batch size were set to 0.0001, 0.05, 50, and 4, respectively. All experiments were conducted on NVIDIA A100 GPUs (80GB).

3.3 Comparison with SOTA Approaches

We compare our DVCTNet with the following typical object detection models: RetinaNet [10], a single-stage detector that addresses class imbalance with focal loss; YOLOX [4], an anchor-free detector that achieves strong performance with a decoupled head; DINO [19], a DETR-like transformer-based detector with denoising training; and Faster RCNN [15], a classical two-stage detector with a region proposal network. We also compare with the recent SOTA detection method FPCL [1], specifically designed for dental caries detection.

Table 1 presents the quantitative comparison results of different object detection methods on both the public dataset [1] and our newly collected benchmark dataset. As evident, our DVCTNet consistently outperforms existing methods across all metrics. Specifically, on the AAAI dataset [1], DVCTNet achieves a slightly better result than the previous SOTA method FPCL by 0.7%, 0.6%, and 1.6%, respectively. The performance gains are more significant on our dual-verified high-quality DVCT dataset, where DVCTNet attains 31.3% AP, 57.4% AP$_{50}$ and 31.9% AP$_{75}$, demonstrating improvements of 14.3%, 14.7% and 21.7% over FPCL and also suppresses the other detection methods at a large scale. Visualization results in Fig. 3 further demonstrate that our DVCTNet can detect tiny caries with low contrast variation across different scenarios. These substan-

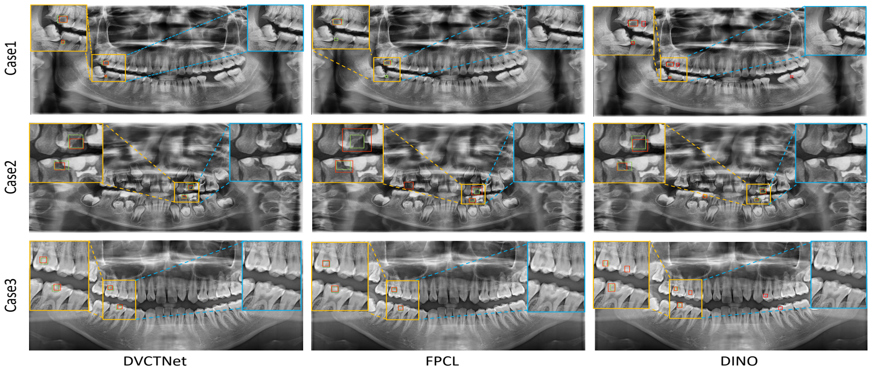

Fig. 3: Visual comparison of dental caries detection methods. Dashed lines mark zoomed views: blue section shows the original image, orange section highlights the same area with overlaid ground-truth (green) and predicted (red) boxes.

Table 2: Ablation analysis for our proposed DVCTNet.

Dual-View Pretraining		**GCV-Atten**	**AAAI Dataset**			**DVCT Dataset**		
Global	**Local**		**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP**	**AP$_{50}$**	**AP$_{75}$**
			32.8	64.3	28.5	15.7	33.0	12.1
✓			45.8	80.1	46.3	27.1	54.0	27.1
✓	✓		46.2	81.7	47.8	27.8	54.6	28.1
✓	✓	✓	**48.9**	**84.7**	**52.2**	**31.3**	**57.4**	**31.9**

tial improvements further highlight the effectiveness of our core idea, which is to develop dual-view co-training that aligns with the clinical diagnosis workflow.

3.4 Ablation Study

We perform ablation experiments to analyze the effectiveness of each key component. As shown in Table 2, The baseline model without any of these components achieves an AP of 32.8% and 15.7% on the AAAI and DVCT datasets, respectively. Adding global-view pretraining significantly improves performance, with AP increasing by 13.0% on AAAI and 11.4% on DVCT. This significant gain demonstrates the importance of foundation model pretraining on panoramic X-rays for caries detection. Incorporating the local-view pretaining model (which is implemented by simply concatenating $\mathbf{f}^{\text{global}}$ and $\mathbf{f}^{\text{local}}$ along the feature channel) enhances the results, with modest gains in AP on the two datasets, suggesting a better dual-view fusion strategy. The addition of GCV-Atten yields the best performance across all metrics, achieving improvements of 2.7% AP, 3.0% AP$_{50}$, and 4.4% AP$_{75}$ on the AAAI dataset, and similar gains on the DVCT dataset. These results validate that each component contributes positively to our framework.

4 Conclusion

In this work, we presented DVCTNet, a novel dual-view co-training framework for accurate dental caries detection that mimics the clinical workflow where dentists combine global panoramic screening with detailed tooth-level inspection. Our DVCTNet effectively integrates the complementary information from both views, leading to superior performance over existing SOTA methods. We also established and released the first dual-verified benchmark dataset from intraoral images and panoramic X-ray images for dental caries detection, named the DVCT dataset, which is the highest-quality dataset for this task so far, with the hope that it will provide a more comprehensive evaluation benchmark for the dental caries detection community.

Acknowledgements. The work was partially supported by National Natural Science Foundation of China (NSFC) under Grant No. 6230012077, Shanghai Municipal Central Guided Local Science and Technology Development Fund under Project No. YDZX20233100001001, Beijing Natural Science Foundation No. L242133, and HPC Platform of ShanghaiTech University.

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

References

1. Chen, B., Fu, S., Liu, Y., Pan, J., Lu, G., Zhang, Z.: Cariesxrays: enhancing caries detection in hospital-scale panoramic dental x-rays via feature pyramid contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21940–21948 (2024)
2. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: The Eleventh International Conference on Learning Representations (2023)
3. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al.: A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. Nature communications **13**(1), 2096 (2022)
4. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
5. Haghanifar, A., Majdabadi, M.M., Haghanifar, S., Choi, Y., Ko, S.B.: Paxnet: Tooth segmentation and dental caries detection in panoramic x-ray using ensemble transfer learning and capsule classifier. Multimedia Tools and Applications **82**(18), 27659–27679 (2023)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016), `https://doi.org/10.1109/CVPR.2016.90`
7. Imak, A., Celebi, A., Siddique, K., Turkoglu, M., Sengur, A., Salam, I.: Dental caries detection using score-based multi-input deep convolutional neural network. Ieee Access **10**, 18320–18329 (2022)

8. Jain, N., Dutt, U., Radenkov, I., Jain, S.: Who's global oral health status report 2022: Actions, discussion and implementation (2024)
9. Karakuş, R., Öziç, M.Ü., Tassoker, M.: Ai-assisted detection of interproximal, occlusal, and secondary caries on bite-wing radiographs: a single-shot deep learning approach. Journal of imaging informatics in medicine pp. 1–14 (2024)
10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (2017)
11. Mei, L., Fang, Y., Cui, Z., Deng, K., Wang, N., He, X., Zhan, Y., Zhou, X., Tonetti, M., Shen, D.: Hc-net: Hybrid classification network for automatic periodontal disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 54–63. Springer (2023)
12. Mohammad-Rahimi, H., Motamedian, S.R., Rohban, M.H., Krois, J., Uribe, S.E., Mahmoudinia, E., Rokhshad, R., Nadimi, M., Schwendicke, F.: Deep learning for caries detection: a systematic review. Journal of Dentistry **122**, 104115 (2022)
13. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal pp. 1–31 (2024)
14. Park, E.Y., Cho, H., Kang, S., Jeong, S., Kim, E.K.: Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. BMC Oral Health **22**(1), 573 (2022)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
16. Wang, X., Guo, J., Zhang, P., Chen, Q., Zhang, Z., Cao, Y., Fu, X., Liu, B.: A deep learning framework with pruning roi proposal for dental caries detection in panoramic x-ray images. In: International Conference on Neural Information Processing. pp. 524–536. Springer (2023)
17. Wen, P., Chen, M., Zhong, Y., Dong, Q., Wong, H.: Global burden and inequality of dental caries, 1990 to 2019. Journal of dental research **101**(4), 392–399 (2022)
18. 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)
19. Zhang, S., Fang, Y., Li, J., Yuan, Z., Sun, L., Wang, C., Wei, Y., Lu, H., Hu, S., Luo, X., et al.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
20. Zhang, X., Liang, Y., Li, W., Liu, C., Gu, D., Sun, W., Miao, L.: Development and evaluation of deep learning for screening dental caries from oral photographs. Oral diseases **28**(1), 173–181 (2022)
21. Zhang, Y., Ye, F., Chen, L., Xu, F., Chen, X., Wu, H., Cao, M., Li, Y., Wang, Y., Huang, X.: Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data **10**(1), 380 (2023)
22. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A survey. Medical Image Analysis p. 103551 (2025)
23. Zhu, H., Cao, Z., Lian, L., Ye, G., Gao, H., Wu, J.: Cariesnet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic x-ray image. Neural Computing and Applications pp. 1–9 (2023)