








This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Prompting Vision-Language Models for Dental Notation Aware Abnormality Detection

Chenlin Du ^{*1} , Xiaoxuan Chen ^{*2}, Jingyi Wang², Junjie Wang³, Zhongsen Li¹ , Zongjiu Zhang^{1,4} , and Qicheng Lao^{3,5}  

¹ School of Biomedical Engineering, Tsinghua University, Beijing, China

² State Key Laboratory of Oral Diseases & National Center for Stomatology & National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, China

³ School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), Beijing, China
qicheng.lao@bupt.edu.cn

⁴ Institute for Hospital Management, Tsinghua University, Beijing, China
zhangzongjiu@mail.tsinghua.edu.cn

⁵ Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract. The large pretrained vision-language models (VLMs) have demonstrated remarkable data efficiency when transferred to the medical domain. However, the successful transfer hinges on the development of effective prompting strategies. Despite progress in this area, the application of VLMs to dentistry, a field characterized by complex, multi-level dental abnormalities and subtle features associated with minor dental issues, remains uncharted territory. To address this, we propose a novel approach for detecting dental abnormalities by prompting VLMs, leveraging the symmetrical structure of the oral cavity and guided by the dental notation system. Our framework consists of two main components: dental notation-aware tooth identification and multi-level dental abnormality detection. Initially, we prompt VLMs with tooth notations for enumerating each tooth to aid subsequent detection. We then initiate a multi-level detection of dental abnormalities with quadrant and tooth codes, prompting global abnormalities across the entire image and local abnormalities on the matched teeth. Our method harmonizes subtle features with global information for local-level abnormality detection. Extensive experiments on the re-annotated DET-NEX dataset demonstrate that our proposed framework significantly improves performance by at least 4.3% mAP and 10.8% AP50 compared to state-of-the-art methods. Code and annotations will be released on <https://github.com/CDchenlin/DentalVLM>.

Keywords: Vision-language models · Dental abnormality detection · Prompting · Dental notation system · Dental panoramic X-ray image.

* Equal contribution.

1 Introduction

Recently, the accomplishments of large-scale pretrained vision-language models (VLMs) such as CLIP [18], GLIP [11], and Grounding DINO [25] have garnered attention. These models undergo initial pretraining to learn universal representations through extensive unlabelled data and have demonstrated data efficiency when transferred to the medical domain [5,13,17,22]. However, the performance of VLMs can be significantly influenced by the prompts used for textual and visual alignment [23], and therefore developing appropriate prompting approaches for medical domain is the fundamental key to the success of transferring [5,17,22].

Despite these advancements, the generalization of VLMs to the field of dentistry remains largely unexplored. In the dental clinical practice, dental panoramic X-ray are universally recognized as fundamental radiography for oral health information [19]. They provide a thorough visualization of all teeth and adjacent structures within a single image, facilitating a preliminary assessment of dental abnormalities, such as dental caries, impacted tooth, and periodontal bone loss [16]. Consequently, an exhaustive analysis of dental panoramic X-ray images is indispensable for screening purposes or therapeutic decision-making. Nonetheless, it has been extensively documented that their interpretation is extremely time-consuming [1,19] and often sensitive to radiologists’ experience [10].

The majority of current machine learning methods developed to assist in interpreting dental panoramic X-ray images are either task-specific [2,9,15,21], or exhibit limitations in accurately localizing fine-grained dental abnormalities [6,12,24]. Notably, none of these methods have harnessed the benefits offered by the dental notation system, e.g., the Fédération Dentaire Internationale (FDI) notation system [20]. The dental notation system can serve as a crucial tool for facilitating precise dental abnormality detection through the identification and record-keeping of individual teeth [4]. Furthermore, it inherently captures the symmetrical structure of the oral cavity, a significant yet under-exploited aspect in dental radiology, which could potentially augment multi-level abnormality detection of dental panoramic X-ray images. Dental abnormalities typically manifest at multiple levels, where some are detectable at a global level across the entire image, such as impacted tooth and residual root, and others are local-level abnormalities, such as dental caries, tooth defect, characterized by the relatively subtle details compared to large dental panoramic X-ray images [15]. These distinctive properties within dentistry could be potentially harnessed to effectively prompt VLMs, thereby enhancing the overall detection of dental abnormalities.

In this paper, we present a dental notation-aware abnormality detection framework by leveraging the dental notation system and incorporating multi-level abnormality prompting. Our framework comprises two main components. Initially, we prompt fine-tuned vision-language models for dental notation-aware tooth detection, which enumerates each tooth to facilitate subsequent dental abnormality detection. Next, our approach embarks on multi-level detection of all fine-grained dental abnormalities. To accomplish this, we first retrieve all the corresponding quadrant codes and then align the tooth codes within the FDI notation system. This strategy is inspired by the clinical practice of dentists where

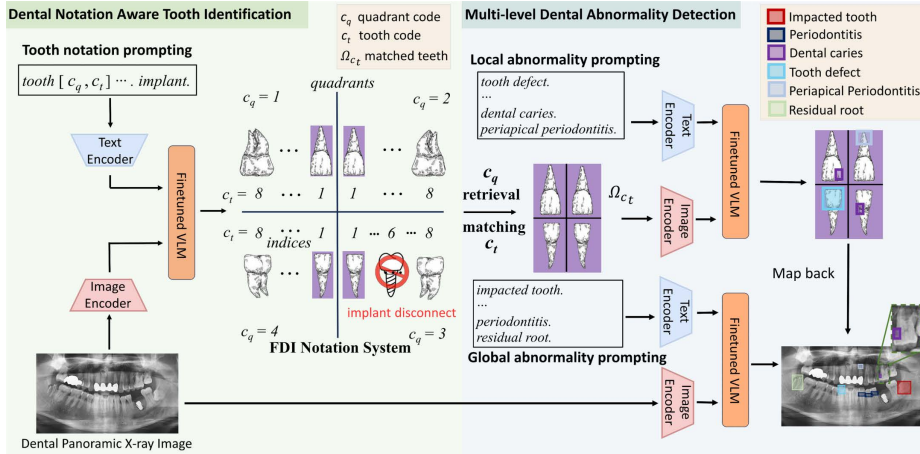


Fig. 1. Illustration of the proposed framework.

analogous teeth in other quadrants are often referred to due to dental symmetry, achieving a balance between the relatively subtle feature and the inclusion of global information and symmetry, particularly in the context of local-level abnormality detection. We then prompt VLMs with global abnormalities across the entire X-ray image and local abnormalities on the matched teeth, respectively, to achieve multi-level dental abnormality detection. Our proposed framework is validated on the re-annotated DENTEX dataset [7], and experimental results demonstrate its superior performance in fine-grained dental abnormality detection, where the overall detection performance is significantly improved by at least 4.3% mAP and 10.8% AP50 compared to state-of-the-art methods.

2 Method

To address the challenge of detecting small dental abnormalities that often present as relatively subtle features within large dental panoramic X-ray images, we propose a robust multi-level detection framework by leveraging a dental notation system, as depicted in Fig. 1 Our framework primarily comprises two components: **dental notation aware tooth identification**, which accurately enumerates each tooth for subsequent detection (Sec. 2.2); and **multi-level dental abnormality detection**, which balances between the relatively small feature map of dental abnormalities and the global information and symmetry available in the image (Sec. 2.3). Together, these components form a comprehensive approach to detecting dental abnormalities within dental panoramic X-ray images.

2.1 Task Formulation

In this paper, we aim to precisely detect all fine-grained dental abnormalities, including impacted teeth, periodontitis, residual roots, tooth defects, den-

tal caries, and periapical periodontitis from dental panoramic X-ray images of those aged 12 years or above. Considering the data scarcity, we employ large pre-trained vision-language models (VLMs), specifically, Grounding DINO[14] in our framework. Given an input dental panoramic X-ray images X , we design text prompt P_A with the class names of the M candidate dental abnormalities, i.e., $P_A = [abnoramlity_1, abnoramlity_2, \dots, abnoramlity_M]$. Therefore, we can predict bounding boxes of candidate abnormal regions $Z_A = [cls, x, y, w, h]$ with $VLM(\cdot, *)$ denoting the prediction process:

$$Z_A = VLM(X, P_A), \quad (1)$$

where cls denotes the class name of the tooth abnormality, while (x, y) is the coordinates of the top-left corner of the box, and (w, h) is the box size. With the following aligning/grounding process, we can fine-tune the pre-trained VLM model with Enc_I and Enc_T representing distinct encoders for image and text prompt, respectively:

$$V = Enc_I(X), U = Enc_T(P_A), S_{ground} = VU^\top, L_{cls} = Loss(S_{ground}, G), \quad (2)$$

where $V \in \mathbb{R}^{N \times d}$, $U \in \mathbb{R}^{M \times d}$ denote the image and text features respectively for N candidate region proposals and M target tooth abnormalities, $S_{ground} \in \mathbb{R}^{N \times M}$ represents the alignment scores, and $G \in \{0, 1\}^{N \times M}$ is the target matrix.

2.2 Dental Notation Aware Tooth Identification

We first argue that conventional dental detection models trained with one-hot labels ignore the symmetry of teeth during tooth enumeration. For instance, despite their anatomical symmetry, central incisors at the upper/lower right and left are usually encoded as $[1, 0, 0, 0, \dots, 0]$, $[0, 1, 0, 0, \dots, 0]$, $[0, 0, 1, 0, \dots, 0]$, $[0, 0, 0, 1, \dots, 0]$, which ignore the semantic and symmetric relationship between teeth. To maximally utilize the symmetry property of teeth for subsequent abnormality detection, we adopt the Fédération Dentaire Internationale (FDI) notation system [20] to locate the tooth T_i by code $C^i = [c_q^i, c_t^i]$ in the dental panoramic X-ray image, where $c_q \in \{1, 2, 3, 4\}$ is the quadrant code, and $c_t \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ is the tooth code. Note that during our preliminary experiments, we found that dental implants might prohibit the accurate matching of tooth notation, thereby deteriorating the subsequent abnormality detection. Thus, we consider detecting dental implants together with tooth identification. The text prompt P_T utilized in this stage can be concluded as follows:

$$P_T = [\text{tooth } C^1, \text{tooth } C^2, \dots, \text{tooth } C^{32}, \text{Implant}]. \quad (3)$$

We then use a fine-tuned VLM to identify and notate all the teeth, while detecting all dental implants in the panoramic image X with prompt P_T :

$$[T, Z_D] = VLM(X, P_T), \quad (4)$$

where $T_i = [C^i, x, y, w, h]$ are the tooth identified at the location C^i in the FDI notation system. Z_D are all detected dental implants, then disconnected.

2.3 Multi-Level Dental Abnormality Detection

To balance the detailed feature map of dental abnormalities with the broader context, we emulate the clinical practice of dentists. Dentists often refer to analogous teeth in other quadrants due to dental symmetry. We achieve this by first retrieving all the corresponding quadrant codes and then matching their tooth codes within the FDI notation system. More specifically, after identifying and numbering all the teeth, each tooth is cropped to form a set of proposed tooth regions, denoted as $\{\tau_{c_q^i, c_t^i}\}$ from the input radiography X . Here $\tau_{c_q^i, c_t^i}$ represents the proposed box region of detected tooth T_i in the image X at the location $[c_q^i, c_t^i]$ in the FDI notation system. For a given set of proposals $\{\tau_{c_q^i, c_t^i}\}$, we retrieve all the c_q and match their c_t .

All the retrieved and matched tooth regions $\{\tau_{1, c_t^i}, \tau_{2, c_t^i}, \tau_{3, c_t^i}, \tau_{4, c_t^i}\}$ are then positioned in a new tooth image $\Omega_{c_t^i}$, according to their c_q . Specifically, regions with c_q from 1 to 4 are placed in the top left, top right, bottom right, and bottom left quadrants, respectively. We conclude this process as the following equation:

$$\Omega_{c_t^i} = \begin{bmatrix} \tau_{1, c_t^i} & \tau_{2, c_t^i} \\ \tau_{4, c_t^i} & \tau_{3, c_t^i} \end{bmatrix}. \quad (5)$$

It is worth noting that we allow absent teeth in the tooth image $\Omega_{c_t^i}$ where a single or multiple $\tau_{c_q^i, c_t^i}$ can be missing.

Subsequently, given all matched tooth images $\{\Omega_{c_t^i}\}$, we can then effectively detect all local-level dental abnormalities by designing prompt with their class names, i.e., $P_{\text{local}} = [\text{tooth defect, dental caries, periapical periodontitis, } \dots]$. We define local-level dental abnormalities as those sensitive to local features, that are more perceivable in the tooth image Ω compared to the original panoramic image X . Predicted results of local abnormalities Z_{local} are then mapped back to the original position in the panoramic image. For global-level abnormalities such as impacted tooth, we directly locate them as Z_{global} from the panoramic image with text prompt $P_{\text{global}} = [\text{impacted tooth, periodontitis, residual root, } \dots]$. Consequently, the multi-level detection of dental abnormalities can be described as follows:

$$Z_A = Z_{\text{local}} \cup Z_{\text{global}} = \text{Map}(VLM(\Omega, P_{\text{local}})) \cup VLM(X, P_{\text{global}}), \quad (6)$$

where $\text{Map}(\cdot)$ denotes the mapping operation of local abnormalities from the tooth image to the original panoramic image.

3 Experiments and Results

3.1 Experimental Setup

Dataset In this study, we employ the quadrant-enumerated subset of the DENTEX dataset [7], encompassing 645 panoramic X-ray images. These X-rays are meticulously annotated with the assistance of dental professionals. It is noteworthy that our annotation process also encompasses dental implants and involves

Table 1. Overview of quantitative comparison with state-of-the-art detection methods for tooth identification, dental implant, and abnormality detection (%).

Method	Tooth		Implant		Abnormality		
	mAP	AP50	AP	AP50	mAP	AP50	
Dyead	[3]	64.5	97.1	49.6	81.6	18.1	36.7
DINO	[25]	63.4	94.3	26.5	48.0	12.6	23.5
SegAndDet	[8]	64.3	96.3	54.7	83.2	20.4	37.7
PDCNN	[9]	60.1	91.6	57.2	93.2	1.2	4.0
HierarchicalDet	[6]	59.8	91.6	54.0	85.6	23.1	46.4
GLIP	[11]	66.7	96.7	71.9	99.6	30.3	52.9
G-DINO	[14]	67.4	97.1	75.2	100	32.7	55.5
Ours		67.4	97.1	75.2	100	37.0	66.3

Table 2. Quantitative comparison with state-of-the-art detection methods for multi-level dental abnormality detection (%).

Method	Local-level Abnormality						Global-level Abnormality						
	Defect		Caries		Periapical		Impacted		Periodontal		Residual		
	AP	AP50	AP	AP50	AP	AP50	AP	AP50	AP	AP50	AP	AP50	
Dyead	[3]	5.9	15.1	1.8	6.6	7.4	23.4	53.1	83.6	35.6	73.4	5.0	18.2
DINO	[25]	3.4	8.6	0.6	2.5	1.3	5.1	46.5	73.4	23.1	48.9	1.0	2.8
SegAndDet	[8]	4.8	11.0	2.7	11.1	8.0	24.6	56.2	76.8	34.6	68.9	16.1	34.1
PDCNN	[9]	0.0	0.0	0.0	0.0	0.0	0.0	2.6	8.8	4.7	15.1	0.0	0.0
HierarchicalDet	[6]	8.6	14.8	5.2	14.8	10.3	30.9	60.5	84.6	32.2	68.0	21.9	65.3
GLIP	[11]	9.1	19.3	6.3	17.4	12.2	36.2	70.5	91.4	39.5	77.8	44.4	75.0
G-DINO	[14]	10.5	21.5	9.8	30.2	16.0	41.7	71.7	91.4	44.9	82.3	43.2	65.9
Ours		21.2	34.0	17.3	48.9	21.6	61.8	71.2	89.8	44.6	82.4	45.9	80.6

the refinement of bounding boxes for each tooth present in every X-ray image. The dataset is randomly partitioned into training, validation, and test sets, containing 405, 102, and 127 panoramic X-ray images. For quantification of the detection performance, we report the average precision (AP) and AP50.

Implementation Details We adopt the Grounding-DINO-T variant [14] as our pre-trained vision-language model. Our models are trained using the AdamW optimizer, with a base learning rate and weight decay both set at 1×10^{-4} . The model with the best performance on the validation set is subsequently utilized for further evaluation. To compensate for potential errors in tooth enumeration, we crop each tooth by an additional 20, 40, 10, and 10 pixels on the crown, root, left, and right sides, respectively. During the inference, all prompts used in our framework were fixed. All baseline models were evaluated with fine tuning using the same data split as our proposed approach.

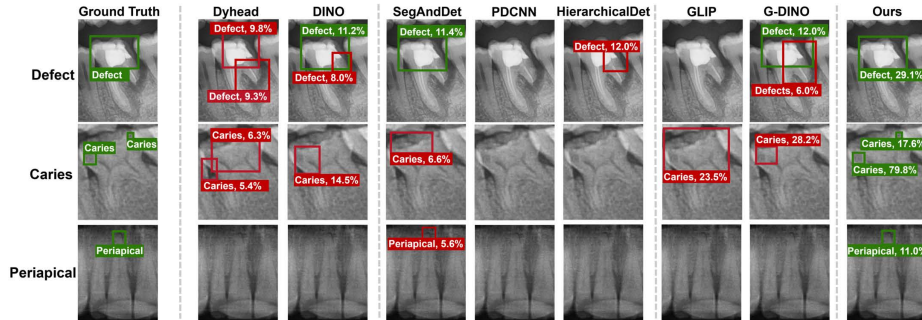


Fig. 2. Visualization of detection results for local-level dental abnormality detection.

3.2 Comparison with State-of-the-Art Methods

To thoroughly evaluate our proposed method, we compare the performance of our framework in tooth identification, dental implant, and abnormality detection against state-of-the-art detection models that can be adapted for dental object detection. These include Dyhead [3], DINO [25], and two VLMs, specifically GLIP [11] and Grounding DINO [14]. Furthermore, our framework has been benchmarked against models specifically designed for dental applications, such as SegAndDet [8], the winner of the DENTEX [7] (MICCAI Challenge 2023), as well as PDCNN [9], and HierarchicalDet [6] to ensure a comprehensive evaluation.

The Proposed Approach Achieves the Best Performance in Dental Abnormality Detection Compared to State-of-the-arts. Table 1 indicates that our framework outperforms all state-of-the-art detection models in dental abnormality detection, as well as tooth and implant identification. Notably, our proposed approach by leveraging the tooth notation prompting exhibits superior performance in dental abnormality detection, obtaining 36.7% mAP and 66.3% AP50 when compared to other baselines where the second highest is Grounding DINO [14] with 32.7% mAP and 55.5% AP50, an increase of 4% in mAP and 10.8% in AP50, respectively. Our results also demonstrate superior accuracy in tooth enumeration and implant detection, especially compared with conventional one-hot encoded models such as Dyhead [3] and DINO [25]. This highlights the effectiveness of our proposed tooth enumeration approach, which leverages the FDI notation system in facilitating subsequent dental abnormality detection.

The Proposed Approach Significantly Improves the Local-level Abnormality Detection. As further elaborated in Table 2, our framework significantly outperforms other methods in detecting local-level dental abnormalities including tooth defect, dental caries, and periapical periodontitis. We observe a significant improvement in both mAP and AP50 scores, with an average increase of 7.9% and 17.2% respectively. The significant improvement underscores the

Table 3. Ablation on key components for local-level detection (%).

Tooth	Notation-aware Implant		Defect		Caries		Periapical		Local-level		
	Cropping	Matching	Disconnection	AP	AP50	AP	AP50	AP	AP50	mAP	AP50
	\times	\times	\times	10.5	21.5	9.8	30.2	16.0	41.7	12.1	31.1
	\checkmark	\times	\times	9.2	22.8	10.2	27.6	11.2	33.7	10.2	28.0
	\checkmark	\checkmark	\times	21.8	36.8	17.1	47.8	18.7	55.6	19.2	46.7
	\checkmark	\checkmark	\checkmark	21.2	34.0	17.3	48.9	21.6	61.8	20.0	48.2

Table 4. Ablation study on designing dental notation prompts (%).

Prompt	Tooth		Implant	
	mAP	AP50	AP	AP50
One-hot	63.4	94.3	26.5	48.0
tooth c_t at quadrant c_q	67.2	96.9	72.7	97.8
tooth $c_q c_t$	67.3	97.0	74.3	99.2

Table 5. Ablation study on multi-level prompting (%).

Detection Level	Abnormality	
	mAP	AP50
Global-level Only	32.7	55.5
Local-level Only	25.8	49.7
Multi-level	37.0	66.3

ability of our framework to effectively amalgamate the detailed features of dental abnormalities with the inherent global and symmetrical information present in dental panoramic X-ray images, facilitated by the dental notation system.

Visualizations Fig. 2 demonstrates our framework’s ability to accurately detect three local-level dental abnormalities. It shows fewer errors in locating tooth defect and superior sensitivity in identifying subtle abnormalities such as dental caries and periapical periodontitis compared to other methods. Full-size dental panoramic X-ray images are available in the appendix.

3.3 Ablation Study

We perform ablation studies to evaluate the effectiveness of the key components in our proposed framework for dental abnormality detection. We first investigate key modules for local-level detection in our dental abnormality detection framework, including tooth cropping, dental notation-aware tooth matching, and implant disconnection. Additionally, we assess the impact of different dental notation prompts on tooth enumeration and implant identification, as well as the influence of multi-level prompting on abnormality detection.

Ablation on key components for local-level detection. Table 3 reveals the indispensability of all three key components proposed for local-level detection. While tooth cropping improves local feature discernment, it compromises the detection of tooth defect and periapical periodontitis due to the loss of global and symmetrical information. This is counterbalanced by the tooth matching, with average performance further boosted by implant disconnection.

Ablation on dental notation prompts. Two FDI notation prompts and one-hot encoding are evaluated for tooth enumeration and implant identification

(Table 4). Both FDI notation prompts outperformed one-hot encoding, with the ‘tooth c_qc_t ’ prompt yielding the best results, more suitable for our framework.

Ablation on multi-level prompting. Table 5 indicates that multi-level prompting, combining global and local prompts, proved effective in addressing the multi-level manifestation of dental abnormalities, compared with global or local only.

4 Conclusion

In this study, we propose a robust, multi-level, dental notation-aware abnormality detection framework that leverages vision-language models to identify fine-grained dental abnormalities. Our methodology, substantiated by comprehensive experimental results, exhibits significant efficacy. Unlike conventional approaches that either scan the entire X-ray image or solely concentrate on isolated tooth images, our framework harnesses the benefits offered by the dental notation system. Our framework adeptly balances the relatively diminutive feature map of dental abnormalities with the globally available information and symmetry. This balance enables it to effectively address the multi-level manifestation of dental abnormalities. A notable constraint of our method is the necessity for multiple fine-tuning stages for each component, which we leave for future work.

Disclosure of Interests. No conflicts of interests to be declared.

References

1. Bruno, M.A., Walker, E.A., Abujudeh, H.H.: Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics* **35**(6), 1668–1676 (2015)
2. Chang, J., Chang, M.F., Angelov, N., Hsu, C.Y., Meng, H.W., Sheng, S., Glick, A., Chang, K., He, Y.R., Lin, Y.B., et al.: Application of deep machine learning for the radiographic diagnosis of periodontitis. *Clinical Oral Investigations* **26**(11), 6629–6637 (2022)
3. Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., Zhang, L.: Dynamic head: Unifying object detection heads with attentions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 7373–7382 (June 2021)
4. Erfan, O., Qasemian, E., Khan, M., et al.: Introduction of new tooth notation systems in comparison with currently in-use systems. *European Journal of Dental and Oral Health* **3**(2), 35–48 (2022)
5. Guo, M., Yi, H., Qin, Z., Wang, H., Men, A., Lao, Q.: Multiple prompt fusion for zero-shot lesion detection using vision-language models. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 283–292. Springer (2023)
6. Hamamci, I.E., Er, S., Simsar, E., Sekuboyina, A., Gundogar, M., Stadlinger, B., Mehl, A., Menze, B.: Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 389–399. Springer (2023)

7. Hamamci, I.E., Er, S., Simsar, E., Yuksel, A.E., Gultekin, S., Ozdemir, S.D., Yang, K., Li, H.B., Pati, S., Stadlinger, B., et al.: Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. arXiv preprint arXiv:2305.19112 (2023)
8. He, L., Liu, Y., Wang, L.: Intergrated segmentation and detection models for dentex challenge 2023. arXiv preprint arXiv:2308.14161 (2023)
9. Kong, Z., Ouyang, H., Cao, Y., Huang, T., Ahn, E., Zhang, M., Liu, H.: Automated periodontitis bone loss diagnosis in panoramic radiographs using a bespoke two-stage detector. *Computers in Biology and Medicine* **152**, 106374 (2023)
10. Kumar, A., Bhadauria, H.S., Singh, A.: Descriptive analysis of dental x-ray images using various practical methods: A review. *PeerJ Computer Science* **7**, e620 (2021)
11. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10965–10975 (2022)
12. Lin, S.Y., Chang, H.Y.: Tooth numbering and condition recognition on dental panoramic radiograph images using cnns. *IEEE Access* **9**, 166008–166026 (2021)
13. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 21152–21164 (October 2023)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
15. 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)
16. de Oliveira Capote, T.S., de Almeida Gonçalves, M., Gonçalves, A., Gonçalves, M.: Panoramic radiography—diagnosis of relevant structures that might compromise oral and general health of the patient. In: *Emerging Trends in Oral Health Sciences and Dentistry*. IntechOpen (2015)
17. Qin, Z., Yi, H.H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. In: *The Eleventh International Conference on Learning Representations* (2022)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
19. Turosz, N., Chęcińska, K., Chęciński, M., Brzozowska, A., Nowak, Z., Sikora, M.: Applications of artificial intelligence in the analysis of dental panoramic radiographs: An overview of systematic reviews. *Dentomaxillofacial Radiology* **52**(7), 20230284 (2023)
20. Van Wijk, A.J., Tan, S.P.: A numeric code for identifying patterns of human tooth agenesis: a new approach. *European journal of oral sciences* **114**(2), 97–101 (2006)
21. 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)

22. Wu, Y., Zhou, Y., Saiyin, J., Wei, B., Lai, M., Shou, J., Fan, Y., Xu, Y.: Zero-shot nuclei detection via visual-language pre-trained models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 693–703. Springer (2023)
23. Yamada, Y., Tang, Y., Yildirim, I.: When are lemons purple? the concept association bias of clip. arXiv preprint arXiv:2212.12043 (2022)
24. 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)
25. 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 (2022)