FD-SOS: Vision-Language Open-Set Detectors for Bone Fenestration and Dehiscence Detection from Intraoral Images

Marawan Elbatel[1], Keyuan Liu[2], Yanqi Yang[2⋆], and Xiaomeng Li[1,3⋆]

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2] Division of Paediatric Dentistry and Orthodontics, Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China
[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China

Abstract. Accurate detection of bone fenestration and dehiscence (FD) is crucial for effective treatment planning in dentistry. While cone-beam computed tomography (CBCT) is the gold standard for evaluating FD, it comes with limitations such as radiation exposure, limited accessibility, and higher cost compared to intraoral images. In intraoral images, dentists face challenges in the differential diagnosis of FD. This paper presents a novel and clinically significant application of FD detection solely from intraoral images. To achieve this, we propose FD-SOS, a novel open-set object detector for FD detection from intraoral images. FD-SOS has two novel components: **conditional contrastive denoising (CCDN)** and **teeth-specific matching assignment (TMA)**. These modules enable FD-SOS to effectively leverage external dental semantics. Experimental results showed that our method outperformed existing detection methods and surpassed dental professionals by 35% recall under the same level of precision. Code is available at: https://github.com/xmed-lab/FD-SOS.

Keywords: Vision Language Model · Open-Set Object Detection

1 Introduction

Bone fenestration and dehiscence (FD) are abnormal conditions that affect the supporting structures of teeth, potentially leading to compromised oral health and tooth loss if left untreated. Achieving accurate detection of FD is crucial for developing effective treatment plans and delivering optimal patient care in dentistry. Currently, dentists primarily rely on analyzing intraoral images of the teeth and gums to make a preliminary diagnosis of FD. However, this diagnostic process often lacks accuracy and suffers from subjectivity. While cone beam computed tomography (CBCT) is considered the gold standard for diagnosing FD

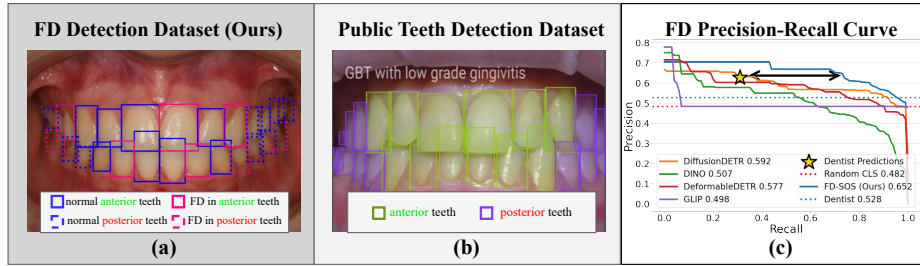⋆ Corresponding authors: eexmli@ust.hk, yangyanq@hku.hk

Fig. 1. (a) Our collected dataset for FD detection focuses on anterior teeth without posterior teeth annotations, as indicated by dashed bounding boxes. (b) Publicly available datasets for teeth detection, incorporating two classes. (c) FD detection Precision-Recall Curve shows that our method outperforms other existing detection methods and even professional dentists. RandomCLS is an always-positive predictor on the true bounding boxes to establish the random average precision (AP).

in clinical practice, its frequent use in orthodontic treatment, especially in pediatric patients, is associated with radiation-related cancer risks [21]. Moreover, CBCT scans are costly and less accessible, particularly in low-income countries. Therefore, there is significant value in developing an FD detection approach that can achieve higher accuracy than dentists while eliminating the need for CBCT scans, using only intraoral images.

While there have been deep learning approaches for detecting FD from CBCT scans [15], no prior work has focused on FD detection from intraoral images. To address this gap, we have collected an in-house dataset called FDTooth, which consists of 150 intraoral images with annotated bounding boxes of anterior teeth provided by dentists, as well as corresponding FD detection results obtained through CBCT. As shown in Fig. 1(a), the dataset only includes annotations for FD and normal cases in anterior teeth. There are no specific annotations for posterior teeth provided, given FD is not visible in frontal view images. This leads to missing annotations for dashed bounding boxes for posterior teeth.

FD detection from intraoral images can be considered an object detection task that involves two subtasks: accurately localizing the teeth ("Anterior" or "Posterior") and diagnosing them as either "No FD" or "FD". One possible solution is to employ object detectors such as DETR [1], YOLO [7], FCOS [20], or Faster-RCNN [17] or their variants like DDETR [25], DINO [23] and Diffusion-DETR [3], with multiple heads and multi-label optimization objectives such as binary cross-entropy, Focal [13], TMLL [9]. Yet, directly applying these methods to our dataset for FD detection yields unsatisfactory results; see "w/o pretraining" in Tab. 2. This can be primarily attributed to the limited size of our dataset, which poses challenges in terms of enlarging it, as our dataset relies on CBCT as the ground truth for each tooth. Therefore, leveraging publicly available teeth detection datasets, which provide detection labels specifically for anterior and posterior teeth (as shown in Fig. 1 (b)), for pretraining the network and subsequently fine-tuning it on our dataset, is a potential solution. Another potential solution is to develop a multi-task object detection framework by using both the

Table 1. Comparisons of our FDTOOTH and Public Teeth Detection dataset from [5].

Feature	FDTOOTH (Ours)	Public Teeth Detection
Image Num.	150	5,000
Resolution	$5,760 \times 3,840$	416×416
BBox Annot.	Normal, FD; 1,800	Anterior, Posterior; 20,000
Disease Label	✓	×
Ground Truth	CBCT scans	None
Train Split	90 imgs (454 healthy, 626 FD)	70% (3,500 imgs)
Val Split	20 imgs (95 healthy, 145 FD)	10% (500 imgs)
Test Split	40 imgs (248 healthy, 232 FD)	20% (1,000 imgs)

public dataset and our dataset to perform "anterior/posterior" teeth detection and "FD/No FD" detection, respectively. However, despite these approaches, we still observe limited results; see "Object Detectors ⋆" in Tab. 2.

Vision-Language Models (VLMs), trained on vast multimodal datasets, have emerged as Open-Set Object Detectors (OSOD), setting new SOTAs in detection across natural image benchmarks [24,12,11,10,14]. For FD detection, fine-tuning OSOD exhibits enhanced performance than standard object detection approaches. Nevertheless, OSOD's effectiveness is constrained, primarily due to catastrophic forgetting and overfitting on limited datasets.

To this end, we present FD-SOS, which stands for **FD S**creening through **O**pen-**S**et object detectors in intraoral images. Specifically, we introduce a **mutli-task VLM** framework that utilizes publicly available dental datasets to enhance FD detection, acting as regularization to mitigate overfitting and improve generalization of VLMs. Given that the public dataset contains a larger number of anterior teeth samples without explicit labels indicating their normal or FD status, we propose a **novel conditional contrastive denoising (CCDN)** to effectively utilize the shared dental semantic information. Moreover, we introduce **teeth-specific matching assignment (TMA)** to enhance the model's confidence in case of missing annotations. Through extensive experiments, our FD-SOS outperform existing detection methods and **surpass dental professionals by 35%** recall under the same level of precision as depicted in Fig. 1(c).

2   Methodology

2.1   Datasets

Since there is a lack of publicly available datasets specifically designed for FD detection from intraoral images, we collected a new dataset called **FDTOOTH**. Table 1 shows a comparison between our dataset and the existing public dataset. Notably, annotations for posterior teeth in our dataset are not available since FD in posterior teeth cannot be detected from a frontal intra-oral view. The public dataset only provides bounding box annotations for all teeth, including both anterior and posterior, without any FD-specific information.
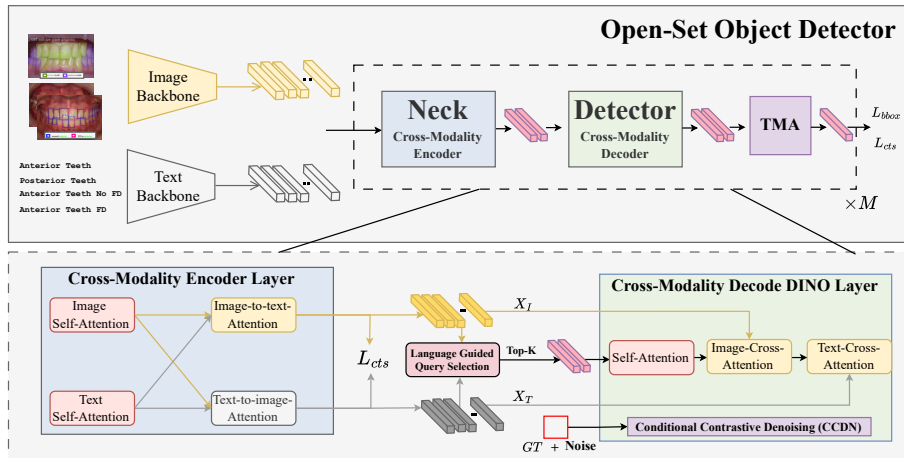
Fig. 2. Framework for FD-SOS.

2.2   Overall Framework

As depicted Fig. 2, we start by extracting text features from each dataset as "Anterior Teeth" and "Posterior Teeth" for teeth detection and "Anterior Teeth No FD" and "Anterior Teeth FD" for FD detection. Leveraging Grounding DINO as our VLM baseline [14], we fuse image-text features at various stages, employing a multimodal image-text contrastive loss to enhance the alignment between the visual and textual modalities. Then, these aligned image-text features are used for object localization and classification tasks. A novel Conditional Contrastive Denoising (CCDN) is proposed during the decoding process for the detection head, DINO [23], exploiting the common semantics of dental structures available in both datasets (e.g "Anterior Teeth"). Finally, we leverage a positional prior to introduce a Teeth-specific Matching Assignment (TMA) to mask out positive detections that are missing in the ground truth.

2.3   Conditional Contrastive Denoising

The ad-hoc approach of encoding the multi-task information in text, although a promising baseline, falls short by not considering the interconnected label space of dental attributes, thus limiting its effectiveness. For instance, G-DINO [14], an OSOD building on DINO [23], employs contrastive denoising to improve training stability. However, its approach to selecting positive and negative anchors for denoising is indiscriminate, overlooking potential spurious correlations caused by overlapping labels across tasks. Specifically, categorizing "FD" in undiagnosed "Anterior teeth" as a negative query fails to recognize that these teeth may be affected by "FD" in a cross-task context. To overcome these challenges, we propose **a novel conditional contrastive denoising (CCDN)** approach.
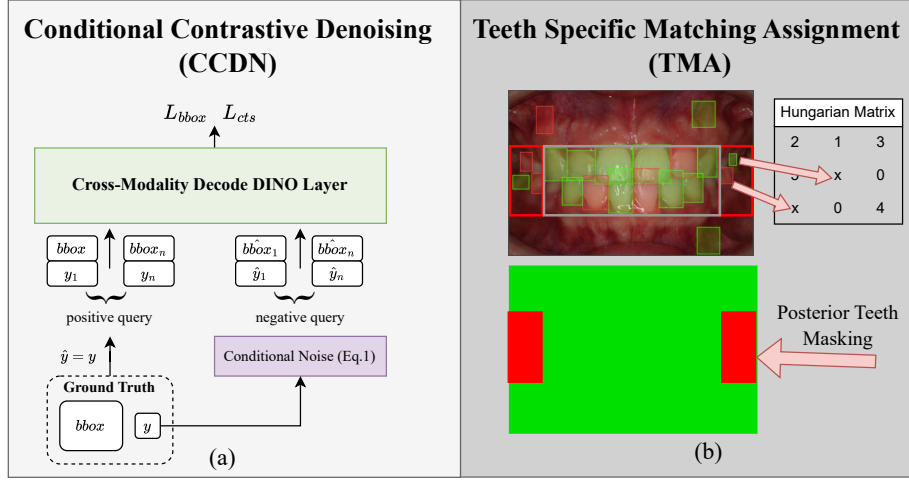
Fig. 3. (a) Conditional Contrastive Denoising (CCDN) improves the detection decoder by utilizing attribute-based contrastive denoising. (b) Teeth-Specific Matching Assignment (TMA) leverages positional priors to mask posterior teeth and focus on the differential diagnosis of FD in frontal intra-oral images.

Let Y denote a random variable representing the query label for the denoising task within DINO head [23]. The label flipping is determined by a probability p conditioned on the dental attributes of the original label y, namely its diagnosis D (where $D = 0$ indicates "No FD" and $D = 1$ indicates "FD"), and position L (where $L = 0$ indicates "Posterior" and $L = 1$ indicates "Anterior"). The conditional probability $P(Y = y' \mid y)$, which defines the likelihood of transitioning from ground truth y to query label y' for the denoising task, is given by:

$$P(Y = y' \mid y) = \begin{cases} 1 - p & \text{if } y' = y, \\ p & \text{if } y' \neq y \text{ and } L_{y'} \neq L_y, \\ \frac{p}{2} & \text{if } y' \neq y \text{ and } L_{y'} = L_y = 1 \text{ and } D_{y'} \neq D_y, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The first condition acts as the **positive query** for reconstruction depicted in Fig. 3 (a). The second condition indicates a position shift (e.g., from "Posterior" to "Anterior" or vice versa) and is considered a **negative query**. When there is a diagnosis available within the "Anterior teeth", the third condition specifies a change by altering the diagnosis, with a probability of $p/2$ to be considered as a **negative query**. The fourth condition denotes cases where transitions do not generate noisy queries due to the uncertainty of adding a diagnosis to an undiagnosed tooth. The bounding box ($BBOX$) noise adheres to the baseline head, DINO [23]. The localization loss, L_{BBOX}, follow DETR-like models [3,23,25] with L1 and Generalized Intersection Over Union (GIOU)

losses. The classification loss leverages contrastive text-image predictions with focal loss [13]. The contrastive classification loss L_{cts} is defined as:

$$L_{cts} = Focal(X_I \cdot X_T^T, y_{GT}) \tag{2}$$

Here X_I and X_T correspond to the image queries and text features respectively. $X_I \cdot X_T^T$ is the dot product output logits, and y_{GT} is ground truth label.

2.4   Teeth-Specific Hungarian Matching Assignment

Object detection models, including OSOD, utilize a sampling approach for the detection head where positive and negative boxes are sampled for a better learning procedure. These models can be characterized by multi-stage [18,17] or single-stage object detectors [20], with the latter being more robust to missing annotations [22]. However, sampling negative queries for training in all object detection models can lead to problems when dealing with imperfect and missing annotations [6,19]. Since the frontal FD detection task is missing annotations within the posterior teeth, it is essential to automatically exclude them during training. To this end, we propose a Teeth-specific Matching Assignment (TMA).

As depicted in Fig. 3 (b). we leverage a **positional prior** about the anterior and posterior teeth, disregarding the posterior teeth during the loss calculation if its information is not available. Utilizing traditional image processing techniques, we identify predicted bounding boxes as posterior teeth if they fall outside the vertical bounds yet lie within the horizontal confines of the anterior teeth region.

Given an image I with a set of bounding boxes $\{(x_1, y_1, x_2, y_2)_b\}$, where $(x_1, y_1)_b$ denotes the top-left corner and $(x_2, y_2)_b$ denotes the bottom-right corner of each bounding box b, we aim to identify the extremities, E, of the anterior teeth region. This can be accomplished by defining the following limits:

$$E_{\text{left}} = \min_b(x_1), \quad E_{\text{right}} = \max_b(x_2), \quad E_{\text{top}} = \min_b(y_1), \quad E_{\text{bottom}} = \max_b(y_2) \tag{3}$$

We then generate a mask for the posterior teeth region as colored in red in Fig. 3 (b) as:

$$mask = ((x_1^p < E_{\text{left}}) \vee (x_2^p > E_{\text{right}})) \wedge ((y_1^p > E_{\text{top}}) \wedge (y_2^p < E_{\text{bottom}})) \tag{4}$$

Following, we assign "x" as a no-care value for predicted bounding boxes within the red mask in the Hungarian matching for object detection [17,1], implicitly masking the unannotated true positives during gradient calculation. This positional prior assignment is simple and effectively handles missing posterior teeth annotations, increasing model confidence without the need for complex architecture or pseudo-label mining, focusing on the differential diagnosis of FD in frontal intra-oral images.

Table 2. Results of FD detection on our dataset. All object detectors commence with initialization from ImageNet pre-trained weights. ⋆ requires pre-training on the publicly dental dataset [5]. † refers to fine-tuning existing VLM pre-trained models. FD-SOS shares the same complexity as the baseline G-DINO (64.02M), higher than DINO (48.04M), and significantly lower than GLIP (122.8M).

Methods	multi-task	AP_{75FD}	AP_{FD}	AP_{50FD}	AP_{75}	AP	AP_{50}
Object Detectors ⋆							
Diffusion-DETR w/o pretraining [3]	×	0.04	1.31	7.58	0.04	1.7	8.85
Diffusion-DETR [3]	×	55.52	51.42	61.28	62.58	59.06	66.37
DDETR [25]	×	56.92	50.41	60.51	62.68	57.44	65.48
DINO [23]	×	54.03	49.68	57.94	55.13	51.65	57.65
Hierarchical-Diff-DetR w/o pretraining [8]	✓	0.0	0.0	0.0	0.0	0.0	0.0
Hierarchical-Diff-DetR [8]	✓	55.81	50.82	59.21	61.45	57.63	64.4
DDETR [25]	✓	53.28	47.57	58.01	60.2	54.89	63.05
DINO [23]	✓	48.12	43.85	50.69	53.8	50.15	55.53
SparseDet [19]	✓	60.5	52.94	62.95	62.96	54.87	65.13
Open-set Object Detectors †							
GLIP [12]	×	40.57	32.0	46.34	51.3	40.47	55.85
GDINO [14]	×	58.32	56.59	61.07	63.69	62.59	65.89
GLIP [12]	✓	41.78	33.68	47.09	51.97	42.73	56.7
GDINO [14] (our baseline)	✓	55.55	54.75	59.99	62.6	62.08	65.81
FD-SOS (ours)	✓	**62.45**	**60.84**	**66.01**	**67.07**	**65.97**	**69.67**

3 Experiments

Implementation Details. For all baselines, we utilize SWIN-T transformer as the image backbone [16], which is initialized with ImageNet weights. The models are trained using AdamW optimizer with a batch size of 16. The learning rate is searched within the range of [2.5e-5,5e-4] with mmdetection framework [2]. For open-set object detectors, we adopt the pre-trained BERT-base as the textual backbone, which was frozen during the fine-tuning process [4].
Baselines. Traditional object detectors include DeformableDETR [25], DINO [23], and DiffusionDet [3] as SOTA object detection models. **Multi-Label Baselines** include Hierarchical-Diff-DETR [8] that utilize DiffusionDet [3]. Additionally, we utilize all SOTA detectors with the cross-task data employing multi-head objectives. **Open Set Object detectors baselines (OSOD)** include GLIP [12] and Grounding DINO [14], SOTA OSOD.
Evaluation Metrics. Metrics include Pascal VOC (AP_{50}) and COCO evaluation (AP_{75}, AP averaged over IoU thresholds from 0.5 to 0.95 with a step of 0.05). We report separate results for the positive FD class, resulting in a total of 6 metrics; For all methods, we exclude posterior teeth model output when calculating metrics on the held-out testing set of **FDTOOTH**. see appendix for teeth detection ("Anterior", "Posterior") on the cross-task dataset.

3.1 Performance on FD Detection

Tab. 2 shows FD detection results on our FDTOOTH test set. A straightforward approach for our task involves applying object detection models without

Table 3. Ablation of proposed components.

	TMA	CCDN	Metrics					
			AP_{75FD}	AP_{FD}	AP_{50FD}	AP_{75}	AP	AP_{50}
GDINO (Baseline)	×	×	55.55	54.75	59.99	62.6	62.08	65.81
w/ TMA	✓	×	59.68	58.29	63.09	64.56	63.63	67.12
w/ CCDN	×	✓	58.59	57.08	61.83	64.34	63.41	66.87
FD-SOS (ours)	✓	✓	**62.45**	**60.84**	**66.01**	**67.07**	**65.97**	**69.67**

multi-task training. Notably, SOTA object detectors without multi-task training and warmup perform poorly, as illustrated by Diffusion-DETR [3], which achieves an AP_{50} of 8.85%, compared to warming up the model, which significantly improves performance to 66.37%. Incorporating cross-task training in the fine-tuning phase, similar to multi-label objectives as seen in Hierarchical-Diff-DETR [8], leads to reduced performance. This decline is probably attributed to spurious correlations and the absence of annotations, with the model being limited by the initial warmup stage's optimum. Consequently, this results in a decrease in Diffusion-DETR's [3] AP_{50} from 66.37% to 64.40%. An alternative approach for missing annotations is employing SOTA sparse-detection solutions from computer vision. Specifically, SparseDet [19] shows enhanced performance achieving AP_{50} of 65.13%, improvement from Hierarchical-Diff-DETR's 64.40%. **Vision-Language Models (VLMs)**, trained on extensive multimodal data and fine-tuned for the FD task without multi-task training, show varied performance. GLIP [12] achieves an AP_{50} of 55.85%, and GDINO [14] achieves 65.89%, reflecting the detrimental effects of spurious correlations present in multi-task frameworks. GLIP struggles with missing annotations due to its use of full-level text attention. In contrast, GDINO's sub-sentence attention mechanism offers an implicit solution to missing annotations, outperforming GLIP and establishing it as a stronger baseline with an improvement of 10% on AP_{50}. Nonetheless, introducing a multi-task framework does not improve GDINO's performance, likely due to spurious correlations and the shared dental semantic and label space. **FD-SOS** address these challenges establishing new SOTA benchmarks across all metrics. Specifically, **FD-SOS** surpass the SOTA object detector Diffusion-DETR [3] by **6.9%** and outperform SparseDet [19] by **11%** on AP, a strict detection metric. Importantly, FD-SOS improves the best-performing **VLM**, GDINO [14], by **3.89%** on the AP metric and **6.09%** on the AP_{FD} metric.

Ablation Study. Tab. 3 shows the ablation results of our FD-SOS. Using our proposed conditional contrastive denoising (CCDN) increases the performance across all metrics, specifically by 2.3% on AP_{FD}. The increase is solely attributed to the removal of spurious correlation during the denoising process, enhancing the performance of the baseline and extending it to multi-task detection. Using only the teeth-specific matching increases the performance by 3.5% on AP_{FD}, indicating that OSOD is still not fully robust to missing annotations, limited by the usage of the detection head, DINO [23]. Utilizing TMA and CCDN together substantially improves the performance for FD by **6.9%** in AP_{75FD}, **6.09%** in AP_{FD}, and **6.02%** in AP_{50FD} compared to the baseline, GDINO [14].
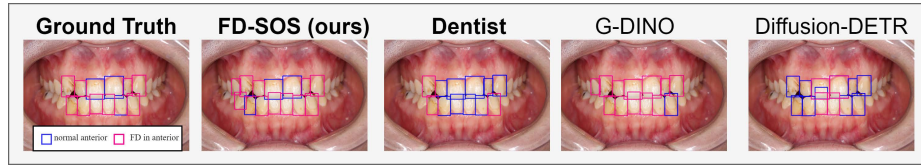
Fig. 4. Qualitative results for the best-performing methods on FD detection.

4   Conclusion

This paper presents a clinically significant application - FD detection from intraoral images. We introduce FD-SOS, a novel OSOD-based framework, which incorporates two innovative components: CCDN and TMA. The results show that our method significantly outperforms existing detection methods and even surpasses the performance of dental professionals, highlighting its immense clinical value in practical settings. Furthermore, this method can be extended to other dental diseases once corresponding datasets become available.

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

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham (2020)
2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19830–19843 (October 2023)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

5. Dwyer, B., Nelson, J., Solawetz, J., et al.: Roboflow (Version 1.0) [Software]. https://roboflow.com (2024), computer vision

6. Elbatel, M., Bornberg, C., Kattel, M., Almar, E., Marrocco, C., Bria, A.: Seamless iterative semi-supervised correction of imperfect labels in microscopy images. In: Kamnitsas, K., Koch, L., Islam, M., Xu, Z., Cardoso, J., Dou, Q., Rieke, N., Tsaftaris, S. (eds.) Domain Adaptation and Representation Transfer. pp. 98–107. Springer Nature Switzerland, Cham (2022)

7. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. ArXiv **abs/2107.08430** (2021), https://api.semanticscholar.org/CorpusID:236088010

8. Hamamci, I.E., Er, S., Simsar, E., Sekuboyina, A., Gundogar, M., Stadlinger, B., Mehl, A., Menze, B.: Diffusion-based hierarchical multi-label object detection toÂ analyze panoramic dental x-rays. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 389–399. Springer Nature Switzerland, Cham (2023)

9. Kobayashi, T.: Two-way multi-label loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7476–7485 (June 2023)

10. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)

11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)

12. Li*, L.H., Zhang*, P., Zhang*, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: CVPR (2022)

13. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2999–3007 (2017), https://api.semanticscholar.org/CorpusID:47252984

14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., yue Li, C., Yang, J., Su, H., Zhu, J.J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. ArXiv **abs/2303.05499** (2023), https://api.semanticscholar.org/CorpusID:257427307

15. Liu, T., Ye, Y., Liu, C., Chen, J., Liu, Y., Xing, W., Ta, D.: Keypoint based automated diagnosis for alveolar dehiscence in mandibular incisors using convolutional neural network. Biomedical Signal Processing and Control **85**, 105082 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2023.105082, https://www.sciencedirect.com/science/article/pii/S1746809423005153

16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

17. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

18. Sun, C.Y., Hong, X.J., Shi, S., Shen, Z.Y., Zhang, H.D., Zhou, L.X.: Cascade faster r-cnn detection for vulnerable plaques in oct images. IEEE Access **9**, 24697–24704 (2021). https://doi.org/10.1109/ACCESS.2021.3056448

19. Suri, S., Rambhatla, S., Chellappa, R., Shrivastava, A.: Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6770–6781 (October 2023)

20. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: A simple and strong anchor-free object detector (2021)

21. Yeh, J.K., Chen, C.H.: Estimated radiation risk of cancer from dental cone-beam computed tomography imaging in orthodontics patients. BMC Oral Health **18** (2018), https://api.semanticscholar.org/CorpusID:51910004

22. Yoon, J., Hong, S., Choi, M.K.: Semi-supervised object detection with sparsely annotated dataset. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 719–723 (2021). https://doi.org/10.1109/ICIP42928.2021.9506790

23. 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 (2023), https://openreview.net/forum?id=3mRwyG5one

24. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. arXiv preprint arXiv:2206.05836 (2022)

25. 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 (2021), https://openreview.net/forum?id=gZ9hCDWe6ke