CariesXrays: Enhancing Caries Detection in Hospital-Scale Panoramic Dental X-rays via Feature Pyramid Contrastive Learning

Bingzhi Chen[1, 3], Sisi Fu[1], Yishu Liu[2*], Jiahui Pan[1], Guangming Lu[2, 3], Zheng Zhang[2*]

[1]South China Normal University, Guangzhou, China
[2]Harbin Institute of Technology, Shenzhen, China
[3]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
{chenbingzhi, fusisi, panjiahui}@m.scnu.edu.cn, {liuyishu, luguangm}@stu.hit.edu.cn, darrenzz219@gmail.com

Abstract

Dental caries has been widely recognized as one of the most prevalent chronic diseases in the field of public health. Despite advancements in automated diagnosis across various medical domains, it remains a substantial challenge for dental caries detection due to its inherent variability and intricacies. To bridge this gap, we release a hospital-scale panoramic dental X-ray benchmark, namely "*CariesXrays*", to facilitate the advancements in high-precision computer-aided diagnosis for dental caries. It comprises 6,000 panoramic dental X-ray images, with a total of 13,783 instances of dental caries, all meticulously annotated by dental professionals. In this paper, we propose a novel *Feature Pyramid Contrastive Learning* (FPCL) framework, that jointly incorporates feature pyramid learning and contrastive learning within a unified diagnostic paradigm for automated dental caries detection. Specifically, a robust dual-directional feature pyramid network (D2D-FPN) is designed to adaptively capture rich and informative contextual information from multi-level feature maps, thus enhancing the generalization ability of caries detection across different scales. Furthermore, our model is augmented with an effective proposals-prototype contrastive regularization learning (P2P-CRL) mechanism, which can flexibly bridge the semantic gaps among diverse dental caries with varying appearances, resulting in high-quality dental caries proposals. Extensive experiments on our newly-established CariesXrays benchmark demonstrate the potential of FPCL to make a significant social impact on caries diagnosis.

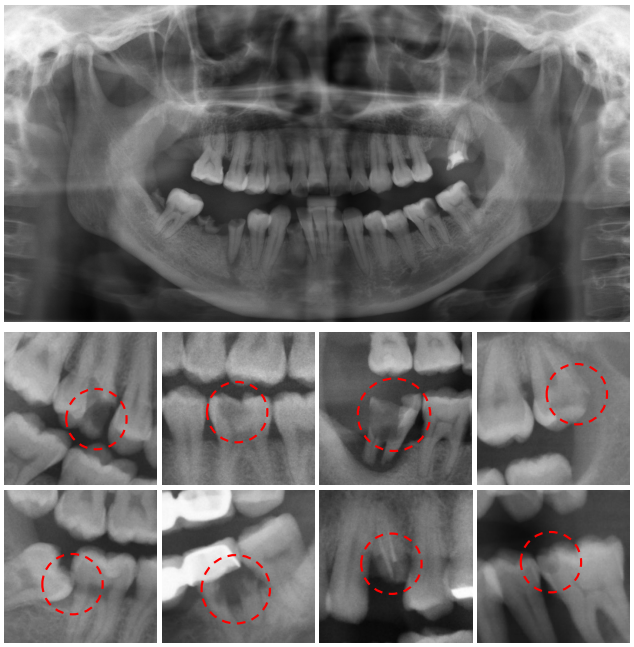


Figure 1: Illustration of panoramic dental X-ray image from our newly-established CariesXrays dataset, which can showcase the diverse variants and appearances of dental caries.

Introduction

Dental caries, known as tooth decay, has become one of the most widespread oral diseases impacting individuals across all age groups (Pitts et al. 2017; Wen et al. 2022). As a commonly utilized imaging technique for diagnosing dental caries in clinical settings, dental panoramic radiography can provide an encompassing overview of the entire dentition structures (Schroder et al. 2019). However, as illustrated in Figure 1, the process of human-visual examinations faces substantial challenges due to many clinical factors, such as diverse lesion appearances, varying sizes, locations, and instances of overlapping structures (Reia et al. 2021). Therefore, the exploration of an AI-driven automated diagnostic paradigm for the screening and detection of dental caries holds a significant social impact on public health, but it still remains largely unexplored and understudied.

With the development of deep learning technologies and the accessibility of digital medical data, the domain of medical image analysis has witnessed significant advancements in various automated diagnosis applications, e.g., breast cancer detection (Wang et al. 2017), pulmonary disease classification (Yan et al. 2018), and brain injury localization (He et al. 2021). However, these advanced techniques heavily rely on the existence of well-annotated medical datasets that typically require expert interpretation and meticulous annotation of the regions of interest. Despite progress made across numerous medical domains, the lack of standardized datasets poses a foundational challenge to researchers working on automated dental caries diagnosis. Technically, the

*Corresponding authors.

task of automated dental caries detection also confronts two critical challenges: 1) *Small targets*: On one hand, caries lesions frequently display subtle alterations characterized by relatively small dimensions in dental radiographs. As a result, conventional image analysis methods encounter a series of challenges in distinguishing between these subtle caries lesions and regular tooth structures. 2) *Multiple variants*: On the other hand, dental caries can manifest in diverse appearances and morphologies, such as radiolucent areas, enamel demineralization, and cavitations (Fraihat et al. 2019). Particularly, the presence of dental restorations, anatomical structures, and artifacts in the radiographs significantly increases the complexity of the detection process.

To address these challenges, we have successfully collected a hospital-scale panoramic dental X-ray benchmark dataset, namely "*CariesXrays*"[1], which is designed to support the advancement of AI-driven automated diagnostic paradigms in the realm of oral health. It comprises 6,000 panoramic dental X-ray images, with a total of 13,783 instances of dental caries. Each instance within the dataset has been meticulously annotated by three dental professionals, ensuring a meticulous and accurate labeling process. Importantly, our CariesXrays dataset comprehensively covers a spectrum of caries presentations, which encompasses a diverse array of forms, morphologies, and stages of progression. In this paper, we also propose a novel Feature Pyramid Contrastive Learning (FPCL) framework that leverages the richness and diversity of CariesXrays to advance the development of automated and accurate dental caries detection. To enhance the model's generalization ability across various caries sizes, a robust dual-directional feature pyramid network (D2D-FPN) is introduced to capture rich and useful contextual information from multi-level feature maps. In addition to the D2D-FPN, we further augment the proposed FPCL model with an effective proposals-prototype contrastive regularization learning (P2P-CRL) mechanism. Through the comparison of model-generated proposals with prototype representations, the P2P-CRL mechanism encourages the mode to capture resilient and discriminative feature representations for various caries variations. Furthermore, a dynamic optimization training strategy with momentum update is applied to enhance the convergence and optimization of the model during the training phase.

As a public health technique, the implementation of automated dental caries detection from panoramic X-ray images carries substantial social impact for improving population oral health and optimizing healthcare resources. It has the potential to revolutionize oral healthcare practices, contributing to the broader advancement of telemedicine and remote healthcare delivery in the field of dentistry. Our main contributions are summarized as follows:

- We propose a large-scale benchmark dataset comprised of panoramic dental X-ray images with high-quality annotations. To our best knowledge, it represents *the first publicly available dataset* towards facilitating the advancement of automated dental caries diagnosis.
- We propose a well-designed Feature Pyramid Contrastive Learning framework that incorporates the advantages of deep learning techniques with the data resources from CariesXrays, to tackle the challenges of the diverse nature of caries manifestations.
- The effectiveness and superiority of our FPCL framework are comprehensively evaluated on CariesXrays. Our approach presents numerous advantages over manual examinations, including increased accuracy, speed, and consistency in identifying carious lesions.

[1]The code and datasource are publicly available at: https://github.com/Binz-Chen/AAAI2024_CariesXrays

Related Work

Medical Image Datasets

The release of large-scale medical image datasets (Wang et al. 2017; Chen et al. 2020a; Yan et al. 2018; Menze et al. 2014; Chen et al. 2023) and the exploration of deep learning techniques have significantly impacted the field of medical image analysis. For instance, the ChestXray dataset (Wang et al. 2017) provides a comprehensive collection of chest X-ray images, allowing researchers to investigate deep learning approaches for the detection and diagnosis of chest diseases. The DeepLesion dataset (Yan et al. 2018) is a comprehensive dataset comprising a diverse collection of annotated lesions extracted from CT scans, enabling researchers to develop and evaluate algorithms for lesion detection and classification. Furthermore, the ISIC archive (Li and Shen 2018) is a collection of dermoscopic images used for the task of melanoma detection and skin lesion analysis, which has played a crucial role in advancing research in computer-aided diagnosis of skin cancers. With the goal of facilitating the development and evaluation of deep learning-based diagnostic models, we endeavor to construct a comprehensive benchmark dataset comprising hospital-scale panoramic dental X-rays to provide a valuable data resource for automated dental caries detection.

Object Detection

Previous studies on object detection have made significant contributions to the advancement of computer vision research. For instance, the introduction of R-CNN (Girshick et al. 2014) is regarded as a groundbreaking milestone to revolutionize object detection by proposing the concept of region proposals, leading to accurate localization and classification of objects within images. Inspired by the accomplishments of R-CNN, subsequent investigations have been driven by the goal of enhancing the speed and efficiency of object detection algorithms, such as Fast R-CNN (Girshick 2015), Faster R-CNN (Ren et al. 2015), and SSD (Liu et al. 2016). Additionally, the incorporation of two-stage detectors, e.g., Mask R-CNN (He et al. 2017) and Cascade R-CNN (Cai and Vasconcelos 2018), has further amplified the capabilities of object detection systems. Furthermore, the YOLO family (Redmon et al. 2016; Wang, Bochkovskiy, and Liao 2023; Ge et al. 2021) has made substantial contributions to the field of computer vision by providing real-time and efficient solutions for object detection. By using a transformer encoder-decoder architecture, DERT (Carion et al. 2020) allows for capturing long-range dependencies

and modeling global contextual information in the object detection process. While making strides in conventional object detection algorithms, the task of dental caries detection remains challenging due to the complex and multifaceted nature of caries presentations.

Contrastive Learning

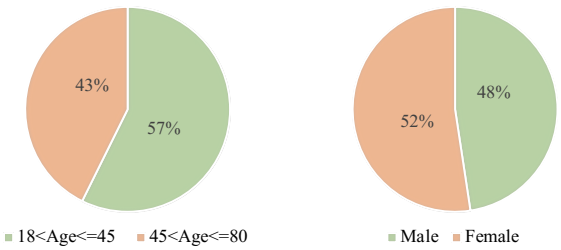
In recent years, contrastive learning (Hadsell, Chopra, and LeCun 2006; He et al. 2020; Chen et al. 2020b) has emerged as a powerful technique for self-supervised representation learning in enhancing feature representation, leading to notable improvements in the performance of object detection methods. For instance, SimCLR (Chen et al. 2020b) effectively learns representations by maximizing agreement between differently augmented views of the same image, which enables the extraction of meaningful and fine-grained features from unlabeled data. To address the inconsistency of the dictionary keys of negative samples, MoCo (He et al. 2020) employs a momentum-based moving average of the query encoder to maintain a large set of negative examples as a source of contrastive supervision. Instead of using negative pairs, BYOL (Grill et al. 2020) employs iterative bootstrapping of network outputs to act as targets for an enhanced representation, which demonstrates the efficacy of contrastive learning in learning powerful image representations. By incorporating multi-level feature supervision and contrastive learning between global images and local patches, DetCo (Xie et al. 2021) proposes a self-supervised contrastive learning approach to effectively enhance the performance of instance-level detection tasks. Inspired by these studies, the technical core of our work is to leverage the capabilities of contrastive learning to enhance feature representation for automated dental caries detection.

Proposed CariesXrays Dataset

In this section, we focus on introducing the proposed CariesXrays dataset within the context of dental caries detection, which serves as the foundation for our research.

Panoramic X-rays Collection

The proposed CariesXrays dataset offers an expansive compilation of panoramic dental X-ray images, which have been meticulously sourced from clinical environments. As depicted in Figure 2, it covers a comprehensive spectrum of variations and patterns exhibited by dental caries across diverse demographic groups and genders, providing a robust representation of the intricate nature of dental caries presentations. To guarantee the quality of data for training and evaluation, a thorough screening process is implemented on the panoramic X-ray films to identify and exclude images with technical issues, such as artifacts, distortions, or improper positioning. Following the previous works (Wang et al. 2017; Johnson et al. 2019), a dedicated preprocessing pipeline is conducted to overcome the challenges of hardware computational capacity. All the images underwent a resizing procedure, resulting in a standardized resolution of 1333×800 pixels as bitmap images. This resizing process ensures compatibility while minimizing any potential loss of essential details crucial for accurate analysis.



(a) Age (b) Gender

Figure 2: Diversity of demographic groups in CariesX-rays.

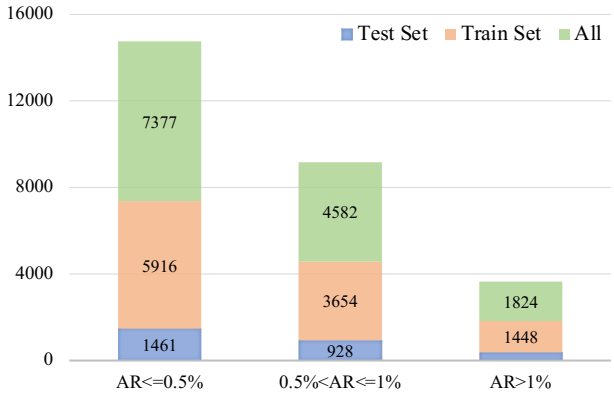


Figure 3: Distribution of labeled objects in CariesX-rays.

Dental Caries Annotations

Technically, the variations observed in panoramic dental radiographs are primarily caused by differences in natural density, thickness, and pathological changes among different parts of the buccal cavity when uniform-intensity X-rays pass through (Sklavos et al. 2019). It is important to note that accurately identifying the boundaries of small caries lesions can be a challenging and time-consuming task. In our study, the annotation procedure within the CariesXrays dataset utilizes a specialized tool to delineate bounding boxes around caries lesions in each image. Inevitably, the complex nature and size variations of caries lesions pose difficulties in precisely delineating their boundaries. To ensure label accuracy, all the annotated outcomes are carefully reviewed by at least three dental professionals.

Dataset Statistics and Distribution

Based on statistical data, the CariesXrays dataset encompasses a total of 6,000 panoramic dental X-ray images obtained from 5,380 distinct patients. Within this dataset, there is a cumulative count of 13,778 meticulously annotated instances of caries lesions. On average, each image in the dental caries anomaly dataset contains approximately 2.3 labeled instances of caries. In our experimental configuration, the dataset is randomly partitioned into two distinct subsets: a training set containing 4,800 images and a test set comprising 1,200 images. Figure 3 illustrates the occupied area ratio (AR) of labeled objects. The predominance of small objects

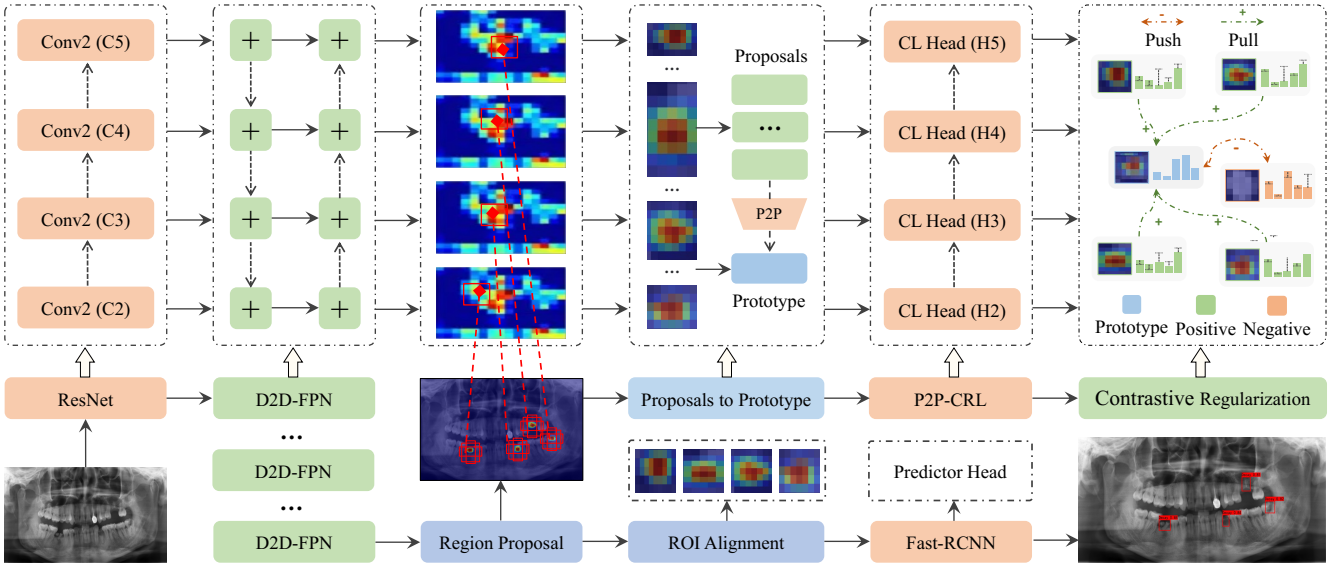


Figure 4: Architecture of the proposed Feature Pyramid Contrastive Learning framework for automated dental caries detection.

(AR $\leq 1\%$) in our dataset highlights the demand for robust and efficient detection algorithms proficient in accurately localizing and identifying these intricate instances.

Proposed FPCL Framework

The main core of the proposed FPCL framework is to optimize the feature extraction process for caries detection while ensuring accurate localization of caries lesions within the panoramic X-ray images. It is built upon the Faster R-CNN architecture (Ren et al. 2015), which has proven to be highly effective in object detection tasks. Figure 4 provides a detailed pipeline of the proposed FPCL framework, comprising two essential modules: 1) dual-directional feature pyramid, and 2) proposals-prototype contrastive regularization.

Dual-Directional Feature Pyramid

Inspired by earlier studies (Liu et al. 2018; Tan, Pang, and Le 2020), we introduce a robust dual-directional feature pyramid network called D2D-FPN. The objective is to efficiently integrate features across diverse scales and resolutions, thereby improving its ability to comprehend both fine-grained details and global contextual information.

Dual-Directional Path In complex panoramic X-ray scenarios, the top-down flow of information may not adequately capture the rich contextual dependencies present in the image. To address this limitation, D2D-FPN is designed to enhance information flow in both top-down and bottom-up directions through a d-layer dual-directional mechanism, facilitating a more comprehensive integration of contextual information. Specifically, we enhance the precise localization signals in the bottom layers by introducing an additional bottom-up pathway, thus creating a fast information pathway between the bottom and top-level features. The bottom-up pathway facilitates the propagation of low-level information,

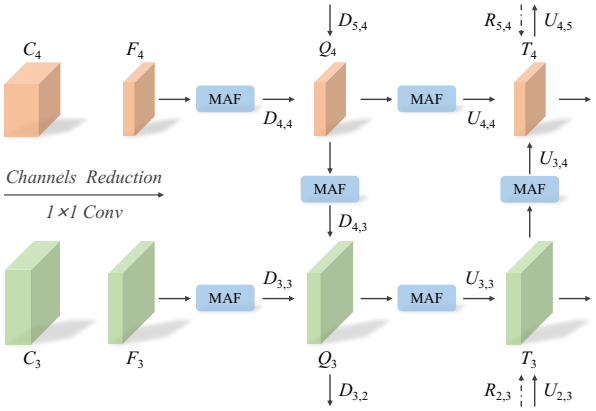


Figure 5: Illustration of the designed D2D-FPN component.

thereby enhancing the model's ability to extract rich contextual information from multiple hierarchical features.

Multi-Scale Fusion Following the previous work (Tan, Pang, and Le 2020), the D2D-FPN module integrates an effective multi-scale attention fusion (MAF) mechanism that can flexibly incorporate learnable weights to capture the importance of different input features. As illustrated in Figure 5, the input features are weighted according to their computed attention weights and subsequently fused to produce the output features, i.e.,

$$\mathcal{Q}_i = \mathcal{D}_{i,i} \cdot F_i + \mathcal{D}_{i+1,i} \cdot \mathcal{Q}_{i+1}, \quad (1)$$

$$\mathcal{T}_i = \mathcal{U}_{i,i} \cdot \mathcal{Q}_i + U_{i-1,i} \cdot \mathcal{T}_{i-1}, \quad (2)$$

where $\mathcal{Q}$ and $\mathcal{T}$ respectively represent the outputs in the top-down and bottom-up streams, $\mathcal{D}$ and $\mathcal{U}$ are the attention weights for fusion. Moreover, our proposed framework incorporates multiple cross-level skip connections with the

weights of $\mathcal{R}$, to facilitate direct information exchange between feature maps originating from different levels.

Objective of Detection Typically, the detection loss in the proposed FPCL method is composed of two main components, i.e., the losses for the RPN stage and the losses for the Fast R-CNN stage, which are designed to capture errors or discrepancies present in both stages of the object detection process (Ren et al. 2015). Mathematically, the detection loss $\mathcal{L}_{\text{Det}}$ used in FPCL can be formulated as,

$$\mathcal{L}_{\text{Det}} = \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RCNN}}, \tag{3}$$

where $\mathcal{L}_{\text{RPN}}$ represents the loss specifically associated with RPN to measure the error in generating accurate region proposals, and $\mathcal{L}_{\text{RCNN}}$ represents the loss associated with Fast R-CNN that focuses on refining the proposals and classifying objects within the proposed regions.

Proposals-Prototype Contrastive Regularization

To address the challenge of insufficient compactness within intra-class distances of dental caries, we propose a proposals-prototype contrastive regularization learning, namely P2P-CRL, which can recalibrate the distance between dental caries proposals and the prototype with the concept of contrastive representation learning.

Proposals to Prototype Due to the inherent variations in the size of proposals, the P2P-CRL module employs subpixel convolution (Shi et al. 2016) to scale all positive proposals to a standard size, without compromising the valuable channel-level information contained within these proposals,

$$\mathcal{X} \to \lfloor \mathcal{X}/r \rfloor, \qquad \mathcal{Y} \to \lfloor \mathcal{Y}/r \rfloor, \tag{4}$$

$$\mathcal{C} \to c \cdot \lfloor r \cdot \text{mod}(\mathcal{Y}, r) + \text{mod}(\mathcal{X}, r) + \mathcal{C} \rfloor, \tag{5}$$

where r denotes the upscaling factor, $\mathcal{X}$ and $\mathcal{Y}$ denote the spatial coordinates within the input proposal, $\mathcal{C}$ denotes the channel index, and c is the number of channels.

In the context of dental caries detection, we aim to project the features of positive proposals and the caries prototype into a latent space, ultimately encouraging proximity between these positive proposals and the class prototype. To achieve this goal, a contrastive loss constraint is applied to measure the similarity between the positive proposals and the class prototype,

$$\mathcal{L}_{\text{CL}} = -\frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} log(< p, \text{Avg}(p) >) + \delta, \tag{6}$$

where $\mathcal{N}$ represents the number of proposal samples, p and $\text{Avg}(p)$ are the normalized representations of the positive proposals and their prototype, $<,>$ refers to a kernel function measuring the similarity between paired vectors, and δ denotes an additional regularization term that used to prevent the negative loss values.

Multi-Level Regularization Our approach prioritizes the application of contrastive learning across multi-level proposals, leading to enhanced performance in detecting dental caries across varying sizes. By incorporating contrastive learning at various levels of granularity, we can comprehensively capture the shared attributes and essential characteristics of caries lesions across various sizes. Hence, the total contrastive loss in P2P-CRL is defined as the sum of the contrastive losses from each level of the CL head, i.e.,

$$\mathcal{L}_{\text{CL}} = \sum \omega_i \cdot \mathcal{L}_{\text{CL}}^i, \tag{7}$$

where ω_i is the hyper-parameter used to balance the contrastive losses for different levels of features.

Momentum-Based Dynamic Optimization

During the training phase, we further propose a momentum-based dynamic optimization training strategy to enhance the model's convergence and optimization. Particularly, as the training advances and the detection loss stabilizes, this strategy dynamically fine-tunes the focus on the contrastive learning task. Based on the foundation of the joint supervision scheme, the training objective of the proposed FPCL approach can be formulated as follows,

$$\mathcal{L}_{\text{Total}}(t) = \mathcal{L}_{\text{Det}}(t) + \lambda(t) \cdot \mathcal{L}_{\text{CL}}(t), \tag{8}$$

where $\lambda(t)$ represents the dynamic weight that is applied to control the contribution of the contrastive loss at each training iteration t. This adaptive adjustment allows the model to gradually shift its focus towards capturing more discriminative information and refining the feature representations.

Momentum Update Inspired by gradient regularization techniques employed in multi-task learning (Athalye, Carlini, and Wagner 2018), we incorporate a momentum update mechanism into our framework, which serves to regulate the extent of influence that the detection loss and contrastive loss have on each other within our proposed method. By introducing this dynamic control mechanism, we aim to strike a balance between the two loss components, optimizing their combined impact on the overall training process,

$$\lambda(t) = \lambda(t-1) + \alpha(\mathcal{L}_{\text{Det}}(t-1) - \mathcal{L}_{\text{Det}}(t)), \tag{9}$$

where α is a smooth factor that represents the degree of influence of detection loss changes on the update of $\lambda(t)$. Consequently, the proposed FPCL model can dynamically adjust and react to evolving patterns within historical training information, enhancing its potential for accurate caries detection and diagnosis across various caries manifestations.

Experiments

Implementation Details

Our FPCL framework is trained on the training set of CariesXRays and evaluated on its validation set. During the training phase, we employ SGD as the parameter optimizer with a batch size of 4 and an initial learning rate of 0.01. The dimension of the feature pyramid channel is set as 256. We select region boxes that simultaneously satisfy a prediction score greater than 0.05 and an intersection over union (IoU) greater than 0.3 as proposals. The final output consists of predicted bounding boxes and confidence scores for identifying areas with dental caries.

Methods	Backbone	Params.	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SSD (Liu et al. 2016)	VGG	34M	12.7	36.2	5.90	-	9.8	24.5
RetinaNet (Lin et al. 2017)	ResNet-50	38M	13.0	30.5	10.2	-	9.4	38.1
DETR (Carion et al. 2020)	Transformer	42M	25.7	64.5	13.7	11.1	23.2	35.0
EfficientDet (Tan, Pang, and Le 2020)	EfficientNet	12M	34.1	52.5	36.0	13.1	39.9	53.7
FCOS (Tian et al. 2019)	ResNet-50	32M	35.9	75.6	29.5	7.5	32.9	49.5
YOLOv7 (Wang, Bochkovskiy, and Liao 2023)	CSPDarkNet	37M	39.3	79.8	34.3	8.8	36.9	50.7
YOLOv8 (Aboah et al. 2023)	CSPDarkNet	11M	40.3	80.7	35.5	8.2	37.7	52.4
YOLOx (Ge et al. 2021)	CSPDarkNet	9M	40.5	81.3	36.1	11.3	37.8	52.1
Conditional-DETR (Meng et al. 2021)	Transformer	43M	42.2	80.6	40.4	18.9	40.8	49.6
Faster R-CNN (Ren et al. 2015)	ResNet-50	41M	39.9	78.0	37.8	9.3	37.8	51.2
FPCL (Ours)	ResNet-50	42M	**48.2**	**84.1**	**50.6**	**18.9**	**47.0**	**55.4**

Table 1: Comparison of Average Precision (%) for automated dental caries detection on our proposed CariesXrays dataset.

Evaluation Metrics

Consistent with the established practices in previous studies (Lin et al. 2017; Ge et al. 2021), our evaluation primarily focuses on reporting the average precision (%) achieved across all benchmark datasets. To make a fair comparison, we utilize the standard Average Precision (AP) metrics under various IoU thresholds, ranging from 0.5 to 0.95.

Baselines

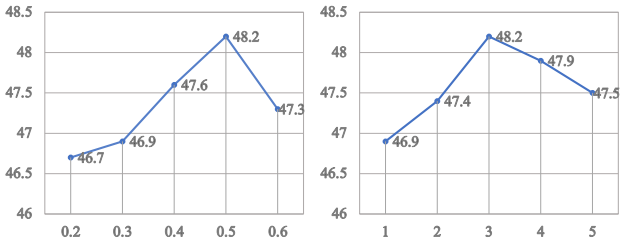
To demonstrate the superiority of the proposed FPCL method, we conduct a comprehensive comparison with a wide range of object detection baselines.

CNN-Based Models The following CNN-based methods are included in the comparison: SSD (Liu et al. 2016), RetinaNet (Lin et al. 2017), EfficientDet (Tan, Pang, and Le 2020), FCOS (Tian et al. 2019), Faster R-CNN (Ren et al. 2015), YOLOv7 (Wang, Bochkovskiy, and Liao 2023), YOLOv8 (Aboah et al. 2023), and YOLOx (Ge et al. 2021).

Transformer-Based Models Additionally, we also compare the proposed FPCL method with various Transformer-based approaches, including DETR (Carion et al. 2020) and Conditional-DETR (Meng et al. 2021).

Comparisons with State-of-The-Art

We evaluate the performance of the proposed FPCL framework through experiments on the CariesXrays dataset compared with the state-of-the-art baselines. The comparison results are presented in Table 1. We can observe that our FPCL method clearly outperforms the comparative baselines in all evaluation metrics, providing more reliable dental caries detection results for oral healthcare applications. Compared with existing CNN-based competitors, such as YOLOv8 and YOLOx, FPCL achieves a significant improvement of 7.7% (48.2% vs. 40.5%) in the AP score on the CariesXrays dataset. Consistently, our FPCL framework exhibits superior performance in handling diverse dental caries presentations, as shown in the detailed metric segmentation for targets of different sizes, surpassing Faster R-CNN in accurately detecting various caries instances. Furthermore, it demonstrates notable improvements over Transformer-based Models, i.e., DETR and Conditional-DETR, with a significant increase of 6.0% (48.2% vs. 42.2%) in AP score. The comparative results presented in our study provide compelling evidence of the effectiveness and robustness of the proposed FPCL method in dental caries detection.



(a) Smooth Factor: α (b) Depth of FPN: d

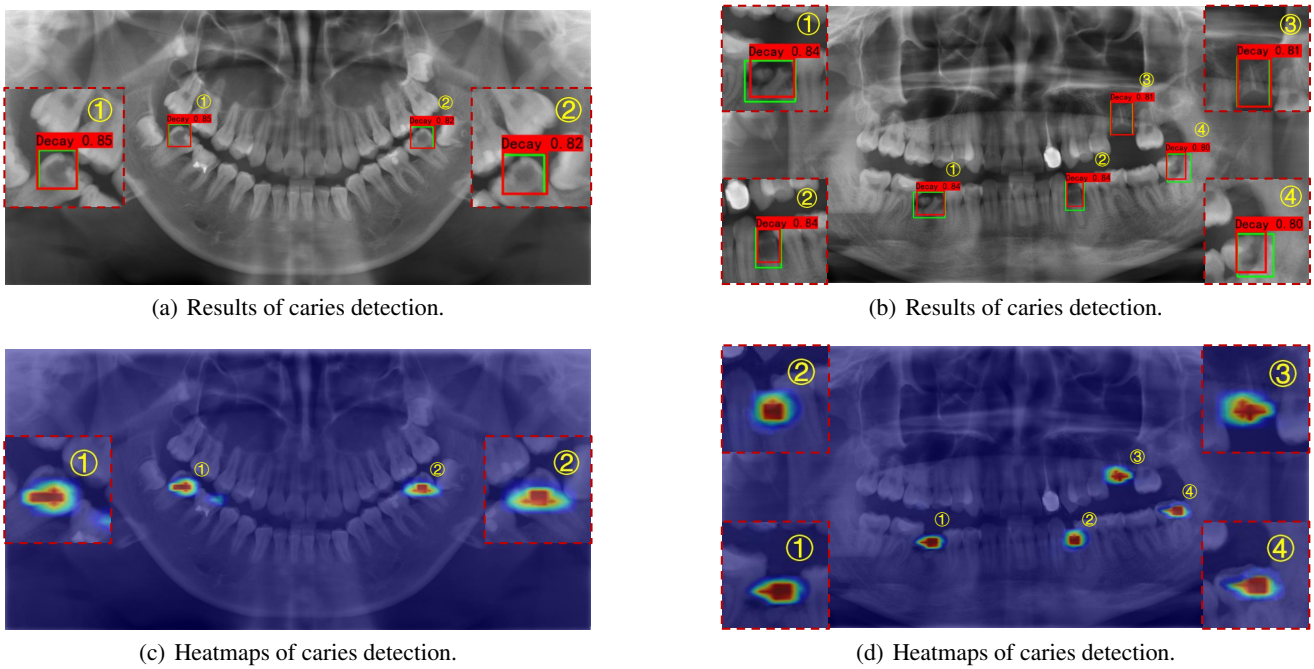
Figure 6: Comparison of AP scores (%) on CariesXrays with different parameter configuration.

Parameter Analysis

In this part, we present an exhaustive parameter analysis of our FPCL method under different parameter configurations. We focus on analyzing the effects of two key parameters: the smooth factor α and the depth of FPN d. The comparative results are summarized in Figure 6.

Evaluation on Smooth Factor The smooth factor α in Eq. 9 is a hyperparameter that controls the balance between the detection and contrastive tasks during training. In our experiments, a smooth factor around 0.5 is found to be most effective in achieving superior performance for the FPCL framework. By contrast, a larger smooth factor may overly emphasize the detection task, which could hinder the model's ability to capture meaningful and discriminative feature representations from training data.

Evaluation on Depth of FPN The depth of FPN refers to the number of D2D-FPN layers used to build the feature pyramid module. From the comparative results, we observe that increasing the depth of FPN from 1 to 3 layers resulted in noticeable performance improvements in terms of AP and accuracy metrics. However, further increasing the depth to 5 layers did not consistently lead to significant improvements.



(a) Results of caries detection.

(b) Results of caries detection.

(c) Heatmaps of caries detection.

(d) Heatmaps of caries detection.

Figure 7: Illustration of bounding box predictions and activation maps generated by the proposed FPCL framework.

Note that deep configurations with more FPN layers could lead to resource-intensive models.

Ablation Studies

In our ablation studies, we perform a thorough analysis of the proposed FPCL method by systematically evaluating the impact of each component on the model's performance. As presented in Table 2, the results demonstrate the importance of each component in FPCL and their collective integration for automated and accurate dental caries detection.

Effect of Feature Pyramid The results of "FPCL w/o FPN" clearly indicate that excluding the proposed D2D-FPN technique leads to reduced robustness, highlighting the crucial role of D2D-FPN in capturing features at various scales and resolutions.

Effect of Contrastive learning Without the contrastive regularization module, "FPCL w/o CRL" shows a notable drop in performance, highlighting the contribution of P2P-CRL in alleviating the semantic gaps among dental caries with varying appearances.

Effect of Dynamic Optimization Consistently, the results for "FPCL w/o DOT" demonstrate that removing the dynamic optimization training strategy substantially decreases the model's performance, due to the model's inability to effectively adapt to the complexity of the training data.

Visualization Results

In Figure 7, we showcase the results of caries detection on two panoramic dental X-ray images using the proposed FPCL model. The images demonstrate the model's ability to accurately identify and localize dental caries regions, represented by the red bounding boxes. It is evident that our FPCL model successfully identifies multiple dental caries regions, providing precise annotations around each affected area. Furthermore, we visualize the attention distribution of feature representations obtained by our model. The results clearly demonstrate the superiority of our FPCL model in capturing meaningful features and exhibiting heightened sensitivity to dental caries regions.

Settings	**FPN**	**CRL**	**DOT**	AP	AP_{50}	AP_{75}
I	✗	✗	✗	39.9	78	37.8
II	✗	✓	✓	47.3	83.5	49.5
III	✓	✗	✓	47.8	83.7	50.3
IV	✓	✓	✗	47.8	83.7	50.6
FPCL	✓	✓	✓	**48.2**	**84.1**	**50.6**

Table 2: Ablation studies (%) for the proposed FPCL method on the CariesXrays dataset.

Conclusion

This paper contributes to broader investigations in the oral health domain by establishing a large-scale panoramic dental X-ray benchmark as well as a well-designed FPCL framework. With the advantages of the feature pyramid network and contrastive learning, the proposed method enables more accurate and automated dental caries detection. Our future work will extend beyond the scope of dental caries detection and focus on a broader range of oral health concerns, potentially revolutionizing dental healthcare practices.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62302172, 62176077), in part by the STI 2030-Major Projects (Grant No. 2022ZD0208900), in part by the Guangdong University Young Innovative Talents Program Project (Grant No. 2023KQNCX020), in part by the Guangdong International Science and Technology Cooperation Project (Grant No. 20220505), in part by the Shenzhen Science and Technology Program (Grant No. RCYX20221008092852077), in part by the Shenzhen Key Technical Project (Grant Nos. 2022N001, 2020N046), in part by the Shenzhen Fundamental Research Fund (Grant No. JCYJ20210324132210025), and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant No. 2022B1212010005).

References

Aboah, A.; Wang, B.; Bagci, U.; and Adu-Gyamfi, Y. 2023. Real-time multi-class helmet violation detection using few-shot data sampling technique and yolov8. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5349–5357. IEEE.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, 274–283. PMLR.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6154–6162. IEEE.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 213–229.

Chen, B.; Li, J.; Lu, G.; Yu, H.; and Zhang, D. 2020a. Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification. *IEEE Journal of Biomedical and Health Informatics*, 24(8): 2292–2302.

Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; and Kong, A. W. K. 2023. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1–14.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1597–1607. PMLR.

Fraihat, N.; Madae'en, S.; Bencze, Z.; Herczeg, A.; and Varga, O. 2019. Clinical effectiveness and cost-effectiveness of oral-health promotion in dental caries prevention among children: systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*, 16(15): 2668.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding yolo series in 2021. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013–2024. IEEE.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1440–1448. IEEE.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 580–587. IEEE.

Grill, J.-B.; Strub, F.; Altche, F.; Tallec, C.; Richemond, P.; and Buchatskaya, E. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 21271–21284. MIT Press.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1735–1742. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738. IEEE.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2961–2969. IEEE.

He, X.; Wang, S.; Chu, X.; Shi, S.; Tang, J.; Liu, X.; Yan, C.; Zhang, J.; and Ding, G. 2021. Automated model design and benchmarking of deep learning models for covid-19 detection with chest ct scans. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 4821–4829. AAAI Press.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317.

Li, Y.; and Shen, L. 2018. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2): 556.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. IEEE.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8759–8768. IEEE.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 21–37.

Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; and Wang, J. 2021. Conditional detr for fast training convergence. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3651–3660. IEEE.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10): 1993–2024.

Pitts, N. B.; Zero, D. T.; Marsh, P. D.; Ekstrand, K.; Weintraub, J. A.; Ramos-Gomez, F.; Tagami, J.; Twetman, S.; Tsakos, G.; and Ismail, A. 2017. Dental caries. *Nature reviews Disease primers*, 3(1): 1–16.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. IEEE.

Reia, V. C. B.; de Toledo Telles-Araujo, G.; Peralta-Mamani, M.; Biancardi, M. R.; Rubira, C. M. F.; and Rubira-Bullen, I. R. F. 2021. Diagnostic accuracy of CBCT compared to panoramic radiography in predicting IAN exposure: a systematic review and meta-analysis. *Clinical Oral Investigations*, 4721–4733.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2164–2173. MIT Press.

Schroder, A. G. D.; de Araujo, C. M.; Guariza-Filho, O.; Flores-Mir, C.; de Luca Canto, G.; and Porporatti, A. L. 2019. Diagnostic accuracy of panoramic radiography in the detection of calcified carotid artery atheroma: a meta-analysis. *Clinical Oral Investigations*, 23: 2021–2040.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1874–1883. IEEE.

Sklavos, A.; Beteramia, D.; Delpachitra, S. N.; and Kumar, R. 2019. The panoramic dental radiograph for emergency physicians. *Emergency Medicine Journal*, 36(9): 565–571.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10781–10790. IEEE.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 9627–9636. IEEE.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7464–7475. IEEE.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2097–2106. IEEE.

Wen, P.; Chen, M.; Zhong, Y.; Dong, Q.; and Wong, H. 2022. Global burden and inequality of dental caries, 1990 to 2019. *Journal of Dental Research*, 101(4): 392–399.

Xie, E.; Ding, J.; Wang, W.; Zhan, X.; Xu, H.; Sun, P.; Li, Z.; and Luo, P. 2021. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 8392–8401. IEEE.

Yan, K.; Wang, X.; Lu, L.; and Summers, R. M. 2018. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3): 036501–036501.