Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# CMViT: A Cross-Modal Pretrained Vision Transformer for Simultaneous Caries and Periapical Lesion Detection in Radiographs

**Chenglin Wang**

wxonet@163.com

State Key Laboratory of Oral Diseases, West China School of Stomatology https://orcid.org/0000-0002-6117-6996

**Ling Ye**

Sichuan University

**Liuyan Huang**

Sichuan University

**Li Shihao**

Department of Biotherapy, West China Hospital, Sichuan University

**Fanyuan Yu**

Sichuan University https://orcid.org/0000-0002-6879-3508

**Feifei Li**

Sichuan University

**Article**

**Keywords:** Self-supervised learning, Vision Transformer, Caries, Periapical lesion, Dual-head

**Posted Date:** December 17th, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-8227244/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

**Additional Declarations:** There is no conflict of interest

## Abstract

The subjective interpretation of periapical radiographs for dental caries and apical periodontitis remains a diagnostic challenge, often failing to detect subtle lesions. To address this, we developed an interpretable deep learning framework that leverages a novel cross-modal self-supervised pretraining strategy on 21 587 unlabeled panoramic single-tooth images, followed by fine-tuning on 6 457 expert-annotated periapical radiographs. Our model employed a dual-head Vision Transformer architecture to simultaneously yet separately classify caries and periapical lesion (PL). This approach significantly enhanced performance, elevating the F1-score from 0.81 to 0.91, with sensitivity rising from 0.76 to 0.90 and specificity from 0.88 to 0.93 compared to training from scratch. The dual-head design also surpassed a single-head, four-class classifier in disease-level accuracy, with notable improvements in complex cases such as teeth with both conditions. Attention rollout heatmaps confirmed that predictions were based on anatomically plausible regions, validating the model's clinical interpretability. Our work demonstrated that cross-modal pretraining combined with a task-specific architecture creates a highly accurate and trustworthy tool for tooth-level diagnosis and quality assurance.


## Introduction

Dental caries and apical periodontitis are among the most prevalent and consequential oral diseases worldwide,[1–4] leading to pain, impaired mastication, and tooth loss with significant public health implications.[5–7] Periapical radiography remains a clinical mainstay for diagnosis, yet its interpretation is inherently subjective, suffering from moderate observer agreement and variable sensitivity due to subtle lesions, anatomical complexity, and human factors like fatigue.[8–12] These limitations motivate the development of automated, objective computer-aided tools for periapical radiographs, which can improve diagnostic consistency, support earlier intervention, and reduce dentists' workload, especially in busy clinics.

Deep learning had already shown promising results on dental radiographs.[13–17] Existing studies had automated tooth-level localization, detection, and classification, which demonstrates the feasibility of artificial intelligence-assisted diagnosis in this domain.[17–19] In routine clinical workflows, clinicians usually first identify individual teeth on the image, either mentally or with software tools. For each tooth, the central question is whether it presents one or more specific pathological conditions, so that its radiographic appearance could be summarized by a small set of diagnostic labels. However, from a modeling perspective, many recent methods therefore focused on learning image representations that support reliable tooth-level labelling on curated datasets, rather than delineating precise lesion boundaries.[20–23] Despite this progress, most existing systems rely on fully supervised training from scratch or initialization from generic ImageNet weights. Such strategies require large annotated datasets and computationally intensive training schedules, and they often show limited robustness when the data distribution shifts across sites, sensors, or acquisition protocols.[19,24,25] In addition, diagnostic labels for each tooth are commonly treated as a single mutually exclusive set, implemented with one softmax classifier. In practice, however, a tooth may exhibit multiple conditions at the same time, such as caries and apical periodontitis. This gap between modeling assumptions and clinical reality complicates supervision, biases evaluation against cases with co-occurring findings, and restricts scalability when new diagnostic categories are introduced.

To address these challenges, we proposed a framework that first leverages self-supervised learning (SSL)[26] to pretrain a Vision Transformer (ViT)[27] backbone on a large-scale dataset of panoramic radiographs,[28] thereby acquiring domain-relevant representations without manual labels and enabling cross-modal transfer to periapical radiographs. Building on this pretrained foundation, we introduced an expandable multi-head architecture that projects shared features into disease-specific subspaces and performs independent binary classification for each condition, thereby natively handling overlapping tooth-level labels and allowing new disease targets to be incorporated with minimal retraining. We empirically shown that this combination of self-supervised cross-modal pretraining and flexible multi-head classification improves caries and apical periodontitis detection on periapical radiographs, and we publicly released the model design and pretrained weights to support future research and facilitate clinically adaptable tools for dental diagnosis.

## Results

We implemented a two-stage study utilizing two distinct datasets (Fig. 1). Initially, we pretrained a Vision Transformer backbone from scratch with Momentum Contrast v3 on the DENTEX panoramic radiograph dataset, using unlabeled single-tooth crops extracted from the panoramic images. Subsequently, we transferred the pretrained encoder to the periapical radiograph task and attached two independent projection heads, one dedicated to caries classification and the other to apical lesion classification, to accommodate overlapping tooth-level labels. The complete model was then fine-tuned on our labeled periapical radiograph dataset using a multi-label objective. To assess the benefit of pretraining, we compared this model with an otherwise identical Vision Transformer trained from scratch under matched settings. We reported sensitivity (SEN), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV), and F1 score. Evaluation was conducted at two levels: a disease-level analysis that reports metrics separately for caries and for apical lesions, and a tooth-level analysis across four clinical categories including control, caries only, apical lesion only, and both. Additionally, we generated attention rollout heatmaps to provide visual explanations of the regions that informed the model predictions.

# Effect of pretraining on ViT performance across disease level and tooth level

Pretraining consistently enhanced performance at the disease level (Table 1). Specifically, pretraining led to an improvement in the overall F1 score from 0.8060 to 0.9146, with sensitivity increased from 0.7594 to 0.9031 and specificity from 0.8750 to 0.9281. Additionally, the positive predictive value rose from 0.8587 to 0.9263, while the negative predictive value improved from 0.7843 to 0.9055. These enhancements indicated a significant reduction in errors, as demonstrated by an increase in true positives from 243 to 289 and true negatives from 280 to 297, alongside a decrease in false negatives from 77 to 31 and false positives from 40 to 23. At the condition level, the F1 score for caries increased from 0.7727 to 0.8994, and for apical lesions, it rose from 0.8407 to 0.9299, primarily driven by sensitivity improvements of approximately fifteen points for caries and fourteen points for apical lesions.

Table 1
Comparison of ViT with and without pretraining at the disease level.

| Model | Condition | TP | FN | TN | FP | SEN | SPEC | PPV | NPV | F1-Score |
|---|---|---|---|---|---|---|---|---|---|---|
| ViT model with pretraining | Caries | 143 | 17 | 145 | 15 | 0.8938 | 0.9063 | 0.9051 | 0.8951 | 0.8994 |
| | PL | 146 | 14 | 152 | 8 | 0.9125 | 0.9500 | 0.9481 | 0.9157 | 0.9299 |
| | Overall | 289 | 31 | 297 | 23 | 0.9031 | 0.9281 | 0.9263 | 0.9055 | 0.9146 |
| ViT model without pretraining | Caries | 119 | 41 | 131 | 29 | 0.7438 | 0.8188 | 0.8041 | 0.7616 | 0.7727 |
| | PL | 124 | 36 | 149 | 11 | 0.7750 | 0.9313 | 0.9185 | 0.8054 | 0.8407 |
| | Overall | 243 | 77 | 280 | 40 | 0.7594 | 0.8750 | 0.8587 | 0.7843 | 0.8060 |

PL: Periapical Lesion

Furthermore, pretraining enhanced tooth-level performance across all four clinical categories (Table 2). Micro F1 increased from 0.6750 to 0.8375 and macro F1 from 0.6727 to 0.8374. Notably, the most substantial class-wise F1 improvements were observed in the co-occurrence category and the apical lesion only category, with increases from 0.6250 to 0.8205 and from 0.6755 to 0.8608, respectively. Caries only improved from 0.6829 to 0.8148 and normal improved from 0.7072 to 0.8537. Sensitivity was higher for every class with pretraining, for example 0.8000 versus 0.5625 in the co-occurrence category and 0.8500 versus 0.6375 in the apical lesion only category. Specificity also increased, with normal rising from 0.8458 to 0.9417. The confusion counts illustrated the same trend, such as a reduction in false positives for normal from 37 to 14 and a reduction in false negatives for the co-occurrence category from 35 to 16.

Table 2
Comparison of ViT with and without pretraining at the tooth level.

| Model | Tooth-level Category | TP | FN | TN | FP | SEN | SPEC | PPV | NPV | F1-Score | Micro-F1 | Macro-F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ViT model with pretraining | Control | 70 | 10 | 226 | 14 | 0.8750 | 0.9417 | 0.8333 | 0.9576 | 0.8537 | 0.8375 | 0.8374 |
| | Caries Only | 66 | 14 | 224 | 16 | 0.8250 | 0.9333 | 0.8049 | 0.9412 | 0.8148 | | |
| | PL Only | 68 | 12 | 230 | 10 | 0.8500 | 0.9583 | 0.8718 | 0.9504 | 0.8608 | | |
| | Both | 64 | 16 | 228 | 12 | 0.8000 | 0.9500 | 0.8421 | 0.9344 | 0.8205 | | |
| ViT model without pretraining | Control | 64 | 16 | 203 | 37 | 0.8000 | 0.8458 | 0.6337 | 0.9269 | 0.7072 | 0.675 | 0.6727 |
| | Caries Only | 56 | 24 | 212 | 28 | 0.7000 | 0.8833 | 0.6667 | 0.8983 | 0.6829 | | |
| | PL Only | 51 | 29 | 220 | 20 | 0.6375 | 0.9167 | 0.7183 | 0.8835 | 0.6755 | | |
| | Both | 45 | 35 | 221 | 19 | 0.5625 | 0.9208 | 0.7031 | 0.8633 | 0.6250 | | |

PL: Periapical Lesion

## Comparison between dual-head structure and sing-head structure using pretrained ViT backbone

Using the same pretrained backbone, the proposed dual-head classifier outperforms the single-head four-class approach when both are evaluated at the disease level (Table 3). Overall F1 improved from 0.8742 with the single head to 0.9146 with the dual head, with sensitivity increasing from 0.8469 to 0.9031 and specificity from 0.9094 to 0.9281. Positive predictive value rose from 0.9033 to 0.9263 and negative predictive value from 0.8559 to 0.9055. The confusion totals shown more correct decisions with the dual head, with true positives increasing from 271 to 289 and true negatives from 291 to 297, while false negatives dropped from 49 to 31 and false positives from 29 to 23.

Table 3
Comparison of dual-head (independent caries and apical-lesion recognition) and single-head (four-class) classifiers using pretrained ViT backbone.

| Recognition method | Condition | TP | FN | TN | FP | SEN | SPEC | PPV | NPV | F1-Score |
|---|---|---|---|---|---|---|---|---|---|---|
| Dual-Head Classifier | Caries | 143 | 17 | 145 | 15 | 0.8938 | 0.9063 | 0.9051 | 0.8951 | 0.8994 |
| | PL | 146 | 14 | 152 | 8 | 0.9125 | 0.9500 | 0.9481 | 0.9157 | 0.9299 |
| | Overall | 289 | 31 | 297 | 23 | 0.9031 | 0.9281 | 0.9263 | 0.9055 | 0.9146 |
| Single-Head Classifier | Caries | 135 | 25 | 145 | 15 | 0.8438 | 0.9063 | 0.9000 | 0.8529 | 0.8710 |
| | PL | 136 | 24 | 146 | 14 | 0.8500 | 0.9125 | 0.9067 | 0.8588 | 0.8774 |
| | Overall | 271 | 49 | 291 | 29 | 0.8469 | 0.9094 | 0.9033 | 0.8559 | 0.8742 |

PL: Periapical Lesion

Per disease, the dual head achieved higher accuracy for both tasks. For caries, F1 increased from 0.8710 to 0.8994, driven by more true positives and fewer false negatives while maintaining the same number of true negatives and false positives. For PL, F1 increased from 0.8774 to 0.9299 with notable improvements in both sensitivity and specificity, supported by an increase in true negatives from 146 to 152 and a decrease in false positives from 14 to 8.

## The confusion matrices for caries and PL detection

The aforementioned findings were further elucidated through the analysis of confusion matrices pertaining to the detection of caries and PL. Comparative panels assessing ViT models with and without pretraining in the binary disease classification task revealed that pretraining significantly improved model performance. Notably, the diagonal elements, indicative of true positives and true negatives, showed substantial increases for both diseases, while off-diagonal errors diminished, thereby signifying enhanced recognition accuracy for caries and PL (Fig. 2a and 2b). In the context of applying the same ViT model to a four-class classification task (Control, Caries only, PL only, and Both), pretraining yielded more precise predictions across all categories. This enhancement was particularly pronounced in the "Both" category, where confusion was significantly reduced, thereby augmenting the model's capacity to differentiate between single and combined lesions (Fig. 2c and 2d). Additionally, utilizing the same pretrained backbone, the confusion matrices facilitated a comparison between the proposed dual-head classifier and the conventional single-head four-class classifier at the disease level. The dual-head classifier exhibited enhanced performance in the detection of both caries and PL, as evidenced by increased true positive and true negative rates, coupled with decreased false positive and false negative rates, resulted in more prominent diagonal entries (Fig. 2e and 2f). These outcomes align with the previously reported higher F1 scores, sensitivity, and specificity, thereby illustrating that the dual-head design surpasses the single-head four-class scheme in disease-level diagnosis.

## Interpretable visual explanations via attention-rollout heatmaps

Initially, we visualized the attention mechanisms associated with carious lesions. The generated heatmaps indicated that the regions with the highest responses were precisely localized to the areas affected by caries (Fig. 3), rather than being diffusely distributed across the entire tooth or the surrounding background. This pattern was consistent across various tooth types, including anterior teeth, premolars, and molars, as well as across different projection angles and image qualities. The high-response areas remained stable and compact, while the other tooth structures and periodontal tissues showed minimal to no activation. Subsequently, we visualized PL. In these instances, the regions of high response were predominantly concentrated around the periapical areas of the roots (Fig. 4), aligning well with the actual locations of the lesions. For multi-rooted teeth with periapical pathology affecting multiple roots, the model effectively highlighted the periapical lesions associated with each affected root. In contrast, the crowns, the middle and coronal portions of the roots, and the surrounding normal alveolar bone were largely unmarked, indicating that the model effectively suppressed background and normal anatomical structures, without generating diffuse high-intensity signals or extensive false-positive regions. Further visualization results of a single tooth affected by both caries and apical periodontitis demonstrated that our imaging modality can simultaneously and distinctly differentiate and localize these two types of lesions within the same tooth. Specifically, the markings near the crown corresponded to the location of caries, while those near the root apex aligned with apical periodontitis (Fig. 5). Additionally, no significant markings were observed in other normal areas of the tooth or the surrounding alveolar bone, indicating the model's effective suppression of background and normal structures, as well as its robust multi-lesion, multi-label recognition capability with minimal mislabeling.

## Discussion

This study established a clinically oriented framework for periapical radiograph analysis by unifying cross-modal self-supervised pretraining on panoramic images with an expandable dual-head Vision Transformer. Our principal finding was that this approach obviously enhanced the concurrent classification of dental caries and PL. Specifically, pretraining the ViT backbone using Momentum Contrast on publicly available panoramic single-tooth crops provided a robust foundational representation, which, when fine-tuned on annotated periapical radiographs, yielded superior performance over models trained from scratch or from generic ImageNet-based initialization.[28,29] The dual-head architecture was pivotal, as replacing a conventional single softmax classifier with independent, disease-specific heads facilitated more adaptable multi-label reasoning that reflected clinical reality, where conditions frequently co-occurred. This design achieved higher F1-scores for

both pathologies by effectively decoupling their predictions. Furthermore, attention-rollout heatmaps confirmed the model's focus on anatomically plausible regions, providing a transparent and educationally valuable justification for its decisions.

Our work extended previous deep learning approaches for dental radiography by offering several advances in how tooth-level disease was modeled and how models were trained. Most existing studies had formulated tooth-level diagnosis as a single-label prediction problem. When both diseases were taken into account, co-existing cases were frequently either aggregated into a single "diseased" category or completely disregarded.[29–32] In such settings, each tooth or tooth surface was assigned exactly one category. Combinations of co-existing diseases were either aggregated into a single label or excluded from analysis. Although this formulation was convenient for model development, it implicitly assumed idealized clinical scenarios. In routine practice, multiple pathologies, such as caries with concomitant apical periodontitis, frequently occur on the same tooth. This discrepancy limited the applicability of such models and reduced their fidelity to real clinical decision-making. From an optimization perspective, many of those systems were trained either from scratch or from generic ImageNet-based initialization.[29,30,33,34] As a result, they typically required larger labeled datasets, longer supervised training schedules, and greater computational and energy resources to reach satisfactory performance. Even so, they still often generalized poorly when deployed in different clinical environments. In contrast, our method leveraged DENTEX, [28] a transformer backbone that had been pretrained with SSL on large-scale panoramic radiographs. This provided a domain-adapted starting point that improved data efficiency. In our experiments, it achieved stronger F1-scores than the scratch baselines we implemented under comparable conditions for multi-label detection of caries and PL. Furthermore, the proposed multi-head architecture was explicitly designed for extensibility. New tooth-level diagnostic targets could be incorporated by attaching additional output heads while retaining the pretrained backbone and previously optimized heads. This design allowed future work to expand functionality without reinitializing or redesigning the entire model. Finally, we released the DENTEX pretrained weights. Researchers could reuse the same backbone for other tooth-centric imaging tasks beyond those studied in our work by loading the weights and fine-tuning on their own annotated datasets. This offered a practical and cost-effective route toward developing a broader family of dental AI models that built on a shared representation.

The architectural innovations of our framework, the ViT's capacity to capture global context and the dual-head classifier's modularity, propel the field toward more flexible and scalable AI tools.[19,35–37] This design not only yielded superior performance but also enables the future integration of additional heads for other dental conditions, such as marginal bone loss or restorations, without retraining the entire backbone.[31,38,39] This aligns with the emerging paradigm of AI systems capable of evaluating a wide array of findings. The model's interpretability and capacity for multi-label detection underscore its potential clinical utility as a secondary reader to enhance diagnostic sensitivity, particularly for subtle lesions in high-workload environments. By providing calibrated, disease-specific probabilities at the tooth level, it can help standardize reporting, facilitate longitudinal monitoring, and support quality assurance and epidemiologic surveillance.[40] The visual explanations afforded by attention maps can further aid dentist-patient communication and trainee education.[17]

Notwithstanding these promising results, several limitations warrant consideration. The computational demands of self-supervised pretraining are significant, and while we optimized under current constraints, greater resources in the future could yield even more robust and stable feature representations. Furthermore, the scarcity of large-scale, public periapical datasets comparable to DENTEX meant we relied on cross-modal transfer to demonstrate feasibility. A systematic comparison with within-modality pretraining on extensive periapical data remains a critical area for future investigation. A systematic comparison in which the encoder is pretrained on a large unlabeled periapical corpus and then fine-tuned on labeled periapical data would be needed to rigorously quantify the gains specific to within-modality self-supervision. These directions will be important to better characterize the true potential and practical limits of the proposed framework.

In conclusion, this framework, by leveraging cross-modal SSL and a flexible dual-head design, consistently outperforms conventional training strategies. Evaluating at the disease level provides feedback more aligned with clinical reasoning than tooth-level summaries, as it respects overlapping conditions. Future work should emphasize model calibration, external multi-site validation, and the exploration of privacy-preserving continual learning methods to ensure the safe and sustainable integration of such tools into clinical practice.

## Materials and methods

# Datasets

In this paper, we used two datasets. The first was the publicly available DENTEX panoramic radiograph dataset (MICCAI 2023; https://dentex.grand-challenge.org). For pretraining, we used only the tooth-level bounding boxes to crop single-tooth regions and treated all crops as unlabeled for SSL. Within DENTEX, 634 panoramic radiographs and 1 005 panoramic radiographs include tooth-level bounding boxes suitable for cropping, while 693 panoramic radiographs provide only quadrant information without tooth boxes and were not used for this step. From the bounding-box subset we obtained 21 587 single-tooth images and used them to pretrain a ViT backbone with Momentum Contrast v3.

The second dataset comprised periapical radiographs from 4 171 patients who underwent dental examinations at West China Hospital of Stomatology between January and December 2017. All teeth visible on each radiograph were cropped to single-tooth images (e.g., incisors, canines, premolars, andmolars). Two board-certified dentists with more than ten years of experience independently assigned one of four labels to each crop: normal, caries only, apical periodontitis only, or caries with apical periodontitis. The readers were blinded to each other's assessments and underwent calibration before annotation; inter-rater agreement exceeded a Cohen's κ of 0.90. Disagreements were resolved by a senior dentist with twenty years of experience, and images without consensus were excluded. This process yielded 6 457 expert-annotated single-tooth images. All images were de-identified prior to analysis, non-diagnostic crops were excluded, and the dataset was partitioned at the patient level into training (6 137 teeth), and testing (320 teeth). All experimental procedures involving human periapical radiographs were approved by the West China Hospital of Stomatology Ethics Committee (WCHSIRB-D-2020-409).

## Dual-Head ViT model with cross-modal pretraining

We pretrain a ViT backbone using a MoCo v3 self-supervised framework on 21 587 unlabeled single-tooth images cropped from the DENTEX panoramic radiographs. Each crop is first zero-padded with black pixels to achieve a 3:4 aspect ratio and then resized to 300 × 400 pixels so that all inputs share a uniform spatial resolution for stable batching and consistent positional encoding. We adopt a MoCo v3 self-supervised framework with a ViT backbone. Each image is encoded into a 384-dimensional class-token embedding, which is then passed through a lightweight projection multilayer perceptron composed of three linear layers to obtain a 256-dimensional vector. During training, two stochastic augmentations of the same image produce paired views. The online pathway yields projected queries q1 and q2 from the projection head, while the momentum pathway provides keys k1 and k2 taken directly from the backbone outputs. The momentum encoder is updated as an exponential moving average of the online encoder to supply stable targets. A symmetric contrastive objective aligns q1 with k2 and q2 with k1 under a temperature-scaled InfoNCE loss with in-batch negatives. All vectors are L2-normalized before similarity computation. This procedure learns view-invariant representations from unlabeled single-tooth images and produces transferable backbone weights for downstream classification.

Following self-supervised pretraining, we prune the MoCo v3 architecture to a single stream. The momentum pathway and its key encoders (k1, k2) are removed entirely. We preserve the parameters of the online ViT encoder and its original three-layer projection MLP, which together map the input image to a 256-dimensional shared representation. These weights initialize the supervised model used for lesion recognition. On top of the shared 256-dimensional latent space, we add two parallel task-specific modules that specialize the representation for caries and for periapical lesions. Each module is a lightweight projection head that takes the shared embedding as input and produces two logits by reducing 256 to 2, followed by a softmax to yield the probability of negative versus positive for the corresponding disease. This design retains a common backbone for feature reuse while allowing disease-specific separation in the final subspaces.

Supervised training uses the second periapical-radiograph dataset. We reuse the same preprocessing as in pretraining to preserve positional encoding fidelity: each single-tooth crop is zero-padded to a 3:4 aspect ratio, resized to 300 × 400 pixels, and normalized. The original four labels are mapped to two binary targets to fit the dual-head formulation. For the caries

head, the target is positive for “caries only” and “caries with apical periodontitis”, and negative for “normal” and “apical periodontitis only”. For the periapical-lesion head, the target is positive for “apical periodontitis only” and “caries with apical periodontitis”, and negative for “normal” and “caries only”. Evaluation was conducted at two levels: a disease-level analysis reporting metrics separately for caries and apical lesions, and a tooth-level analysis across four clinical categories (control, caries only, apical lesion only, and both), supplemented by attention rollout heatmaps that visualize the regions informing each prediction.

Optimization is end-to-end over the backbone, the shared projection MLP, and both disease heads. For each image, we compute a standard cross-entropy loss on the caries head and a separate cross-entropy loss on the periapical-lesion head, where each head predicts negative or positive for its task. The total objective is the sum of the two losses, $\mathcal{L}_{total} = \mathcal{L}_{caries} + \mathcal{L}_{periapical}$, and a single backward pass updates all trainable parameters. At inference, the two softmax outputs provide independent probabilities for caries and for periapical lesions, enabling recognition of normal teeth, single-disease cases, and co-occurrence when both heads are positive (Fig. 6).

Statistical Metrics

Model performance was quantified using SEN, SPEC, PPV, NPV, and F1 score. For a binary target with true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), we compute

$$SEN = \frac{TP}{TP + FN}$$

1

$$SPEC = \frac{TN}{FP + TN}$$

2

$$PPV = \frac{TP}{TP + FP}$$

3

$$NPV = \frac{TN}{FN + TN}$$

4

$$F_1 score = \frac{2 * PPV * SEN}{PPV + SEN}$$

5

At the disease level, these metrics are computed separately for the caries head and for the apical-lesion head using their respective binary labels. Consistent with our label mapping, the caries head is positive for “caries only” and “caries with apical lesion,” and negative for “normal” and “apical lesion only.” The apical-lesion head is positive for “apical lesion only” and “caries with apical lesion,” and negative for “normal” and “caries only.” Probabilities from each head are converted to class decisions using a decision threshold of 0.5.

At the tooth level, evaluation follows the four clinical categories of control, caries only, apical lesion only, and both. We report SEN, SPEC, PPV, NPV, and F1 for each class under a one-versus-rest formulation, yielding class-wise metrics that reflect retrieval of each clinical phenotype against the remaining categories. To provide aggregate summaries across classes, we

additionally report macro-averaged scores, computed as the unweighted mean of the per-class values, and micro-averaged scores, computed by pooling TP, TN, FP and FN across all classes before applying the formulas above.

# Declarations

### Data availability

The source code implementing the dual-head Vision Transformer model, along with pretrained weights, is publicly available on Zenodo at DOI: 10.5281/zenodo.17627584.


### Acknowledgements

This work was supported by Technology Program 2024JDKXJ0001 (L.Y.), Exploration, Research Projects of West China Hospital of Stomatology LCYJ2019-18 (L.Y.), Sichuan Provincial Engineering Research Center of Smart Diagnosis and Treatment for Oral Diseases 30077 (C.L. W.), the Sichuan University, Natural Science Foundation of Sichuan Province 2024NSFSC0551(C.L. W.), and Sichuan Science and Technology Program 2025ZNSFSC1592 (S.H. L.)


### Declaration of Interests

The authors declare no conflict of interest.

### Author Contributions

Liuyan Huang, contributed to conception, design, data acquisition and interpretation, and drafted manuscript. Shihao Li, contributed to conception, design, data acquisition and interpretation, and drafted manuscript. Fanyuan Yu, contributed to data acquisition and critically revised the manuscript. Feifei Li, contributed to data acquisition and critically revised the manuscript. Chenglin Wang, Ling Ye, contributed to design, data interpretation, and critically revised the manuscript.

# References


1. Arias, Z. et al. Recent Advances in Apical Periodontitis Treatment: A Narrative Review. Bioengineering (Basel, Switzerland) 10, doi:10.3390/bioengineering10040488 (2023).
2. Yactayo-Alburquerque, M. T., Alen-Méndez, M. L., Azañedo, D., Comandé, D. & Hernández-Vásquez, A. Impact of oral diseases on oral health-related quality of life: A systematic review of studies conducted in Latin America and the Caribbean. PloS one 16, e0252578, doi:10.1371/journal.pone.0252578 (2021).
3. Warreth, A. Dental Caries and Its Management. International journal of dentistry 2023, 9365845, doi:10.1155/2023/9365845 (2023).
4. Pohl, S. et al. Understanding dental pulp inflammation: from signaling to structure. Frontiers in immunology 15, 1474466, doi:10.3389/fimmu.2024.1474466 (2024).
5. Bramantoro, T. et al. Early Childhood Caries, Masticatory Function, Child Early Cognitive, and Psychomotor Development: A Narrative Review. European journal of dentistry 18, 441-447, doi:10.1055/s-0043-1774326 (2024).
6. Cheng, L. et al. Expert consensus on dental caries management. Int J Oral Sci 14, 17, doi:10.1038/s41368-022-00167-3 (2022).
7. Kassebaum, N. J. et al. Global, Regional, and National Prevalence, Incidence, and Disability-Adjusted Life Years for Oral Conditions for 195 Countries, 1990-2015: A Systematic Analysis for the Global Burden of Diseases, Injuries, and Risk Factors. J Dent Res 96, 380-387, doi:10.1177/0022034517693566 (2017).
8. Meusburger, T. et al. The Detection of Dental Pathologies on Periapical Radiographs-Results from a Reliability Study. Journal of clinical medicine 12, doi:10.3390/jcm12062224 (2023).

9. Siddique, S. N. et al. Quality Assessment of Periapical Radiographs Taken by Dental Assistants Using the Recent Faculty of General Dental Practice (FGDP) Guidelines. Cureus 16, e68508, doi:10.7759/cureus.68508 (2024).
10. Weber, L. S. & Schulze, R. The effects of examiner fatigue on the diagnostic accuracy of dental radiographs. Clinical oral investigations 25, 6193-6199, doi:10.1007/s00784-021-03918-4 (2021).
11. Lam, E. W. N. et al. Interexaminer Agreement in the Radiologic Identification of Apical Periodontitis/Rarefying Osteitis in the National Dental Practice-Based Research Network PREDICT Endodontic Study. Journal of endodontics 47, 1575-1582, doi:10.1016/j.joen.2021.07.008 (2021).
12. Hegde, S., Gao, J., Vasa, R., Nanayakkara, S. & Cox, S. Australian Dentist's Knowledge and Perceptions of Factors Affecting Radiographic Interpretation. International dental journal 74, 589-596, doi:10.1016/j.identj.2023.11.006 (2024).
13. Heo, M. S. et al. Artificial intelligence in oral and maxillofacial radiology: what is currently possible? Dento maxillo facial radiology 50, 20200375, doi:10.1259/dmfr.20200375 (2021).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436-444, doi:10.1038/nature14539 (2015).
15. Ekert, T. et al. Deep Learning for the Radiographic Detection of Apical Lesions. J Endod 45, doi:10.1016/j.joen.2019.03.016 (2019).
16. Schwendicke, F. et al. Cost-effectiveness of Artificial Intelligence for Proximal Caries Detection. J Dent Res 100, 369-376, doi:10.1177/0022034520972335 (2020).
17. Schwendicke, F., Samek, W. & Krois, J. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res 99, 769-774, doi:10.1177/0022034520915714 (2020).
18. Liu, J. et al. Periapical lesion detection in periapical radiographs using the latest convolutional neural network ConvNeXt and its integrated models. Scientific reports 14, 25429, doi:10.1038/s41598-024-75748-9 (2024).
19. Felsch, M. et al. Detection and localization of caries and hypomineralization on dental photographs with a vision transformer model. NPJ Digit Med 6, 198, doi:10.1038/s41746-023-00944-2 (2023).
20. Hamdan, M. H. et al. The effect of a deep-learning tool on dentists' performances in detecting apical radiolucencies on periapical radiographs. Dento maxillo facial radiology 51, 20220122, doi:10.1259/dmfr.20220122 (2022).
21. Issa, J. et al. Diagnostic Test Accuracy of Artificial Intelligence in Detecting Periapical Periodontitis on Two-Dimensional Radiographs: A Retrospective Study and Literature Review. Medicina (Kaunas, Lithuania) 59, doi:10.3390/medicina59040768 (2023).
22. Wang, K. X. et al. Deep learning-based efficient diagnosis of periapical diseases with dental X-rays. Image Vis. Comput. 147, 11, doi:10.1016/j.imavis.2024.105061 (2024).
23. Shafı, I. et al. Teeth Lesion Detection Using Deep Learning and the Internet of Things Post-COVID-19. Sensors (Basel, Switzerland) 23, doi:10.3390/s23156837 (2023).
24. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. Nat Commun 11, 3673, doi:10.1038/s41467-020-17478-w (2020).
25. Roschewitz, M. et al. Automatic correction of performance drift under acquisition shift in medical image classification. Nat Commun 14, 6608, doi:10.1038/s41467-023-42396-y (2023).
26. Huang, S.-C. et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med 6, 74, doi:10.1038/s41746-023-00811-0 (2023).
27. Azad, R. et al. Advances in medical image analysis with vision Transformers: A comprehensive review. Medical image analysis 91, 103000, doi:10.1016/j.media.2023.103000 (2024).
28. Hamamci, I. E. et al. DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Diagnosis Benchmark for Panoramic X-rays arXiv. arXiv (2023).
29. Li, S. et al. Artificial intelligence for caries and periapical periodontitis detection. J Dent 122, 104107, doi:10.1016/j.jdent.2022.104107 (2022).

30. da Silva-Filho, J. E. et al. Deep learning for detecting periapical bone rarefaction in panoramic radiographs: a systematic review and critical assessment. Dentomaxillofac Radiol 54, 405-419, doi:10.1093/dmfr/twaf044 (2025).
31. Ibraheem, W. I. et al. Assessment of the Diagnostic Accuracy of Artificial Intelligence Software in Identifying Common Periodontal and Restorative Dental Conditions (Marginal Bone Loss, Periapical Lesion, Crown, Restoration, Dental Caries) in Intraoral Periapical Radiographs. Diagnostics (Basel) 15, doi:10.3390/diagnostics15111432 (2025).
32. Liu, J. et al. A comparative analysis of deep learning models for assisting in the diagnosis of periapical lesions in periapical radiographs. BMC Oral Health 25, 801, doi:10.1186/s12903-025-06104-0 (2025).
33. Çelik, B., Savaştaer, E. F., Kaya, H. I. & Çelik, M. E. The role of deep learning for periapical lesion detection on panoramic radiographs. Dentomaxillofac Radiol 52, 20230118, doi:10.1259/dmfr.20230118 (2023).
34. Chen, I. D. S. et al. Deep Learning-Based Recognition of Periodontitis and Dental Caries in Dental X-ray Images. Bioengineering (Basel) 10, doi:10.3390/bioengineering10080911 (2023).
35. van Nistelrooij, N. et al. Automated detection and classification of osteolytic lesions in panoramic radiographs using CNNs and vision transformers. BMC Oral Health 25, 950, doi:10.1186/s12903-025-06209-6 (2025).
36. Dujic, H. et al. Automatized Detection of Periodontal Bone Loss on Periapical Radiographs by Vision Transformer Networks. Diagnostics (Basel) 13, doi:10.3390/diagnostics13233562 (2023).
37. Vinayahalingam, S. et al. Advancements in diagnosing oral potentially malignant disorders: leveraging Vision transformers for multi-class detection. Clin Oral Investig 28, 364, doi:10.1007/s00784-024-05762-8 (2024).
38. Shan, T., Tay, F. R. & Gu, L. Application of Artificial Intelligence in Dentistry. J Dent Res 100, 232-244, doi:10.1177/0022034520969115 (2020).
39. Kühnisch, J., Meyer, O., Hesenius, M., Hickel, R. & Gruhn, V. Caries Detection on Intraoral Images Using Artificial Intelligence. J Dent Res 101, 158-165, doi:10.1177/00220345211032524 (2021).
40. Wen, P. Y. F., Chen, M. X., Zhong, Y. J., Dong, Q. Q. & Wong, H. M. Global Burden and Inequality of Dental Caries, 1990 to 2019. J Dent Res 101, 392-399, doi:10.1177/00220345211056247 (2021).

## Figures

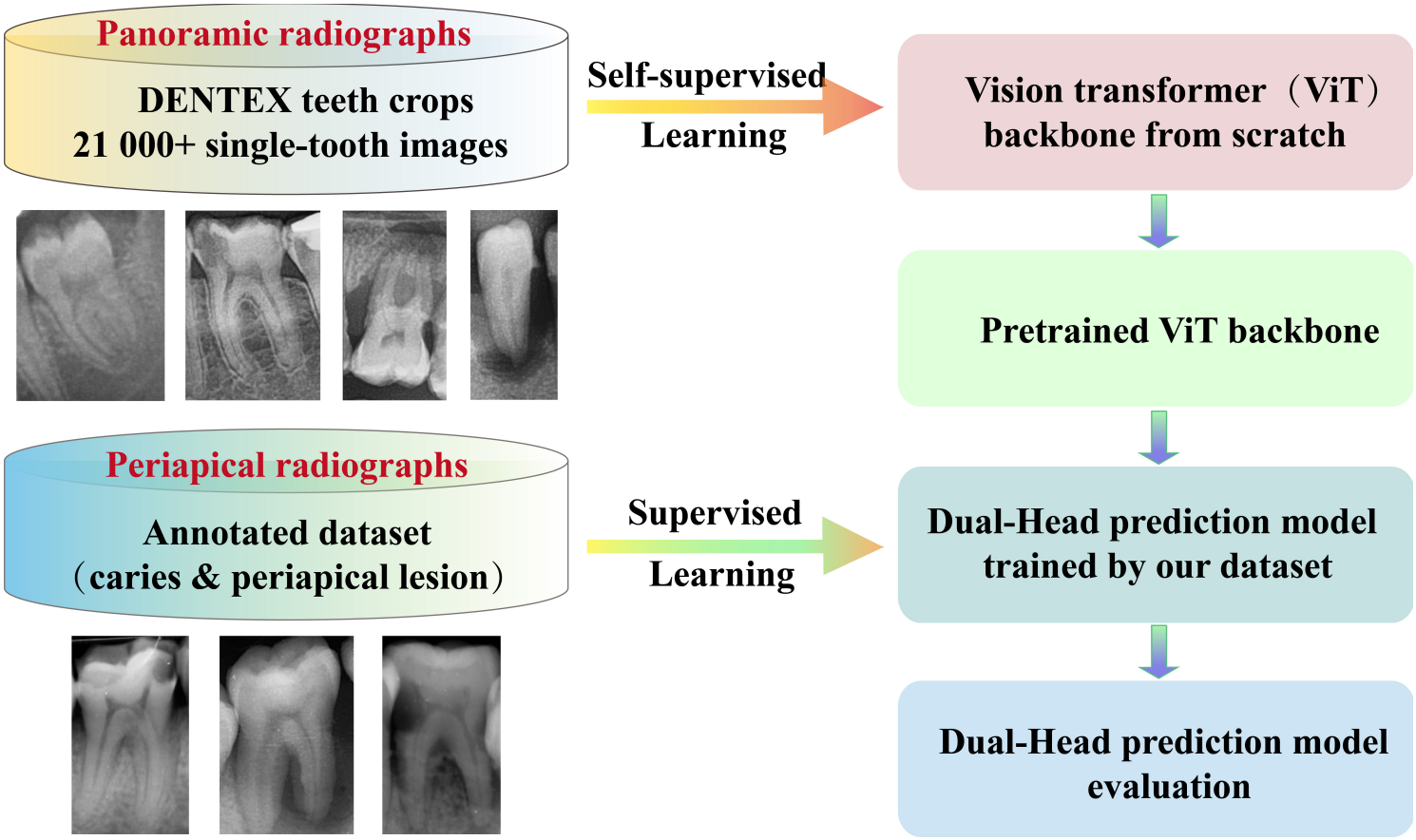


**Figure 1**

Workflow of the proposed model for CMViT disease prediction. The model first utilizes panoramic radiographs from the DENTEX dataset containing over 21 000+ single-tooth images for SSL to pretrain the ViT backbone. After pretraining, the ViT backbone is fine-tuned on periapical radiographs with an annotated dataset for caries and apical periodontitis using supervised learning. The final dual-head prediction model is then evaluated for its performance in detecting both diseases on periapical radiographs.

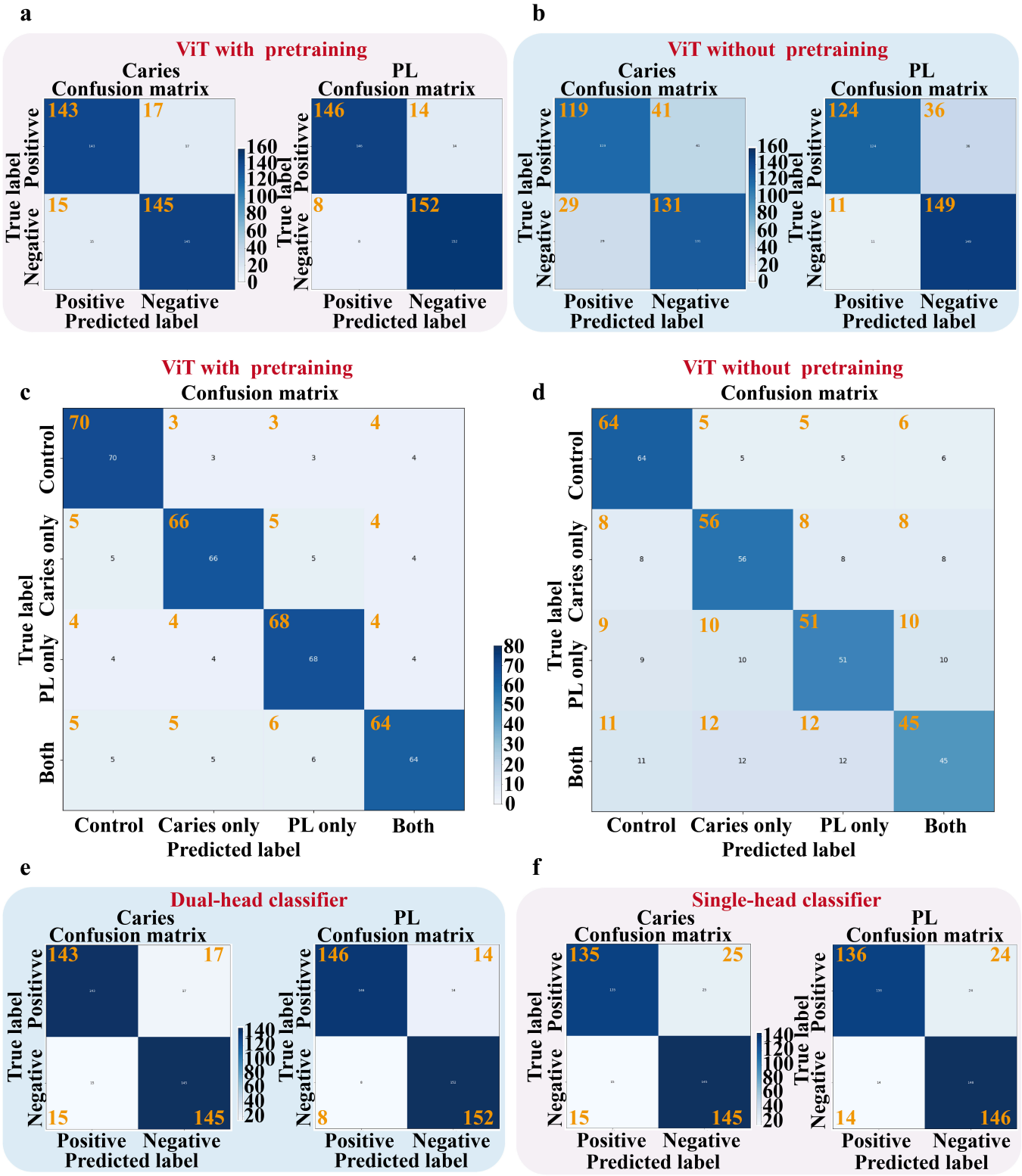


**Figure 2**

Confusion matrices of ViT-based models under different training strategies and classifier designs. (a, b) Disease-level confusion matrices for caries and periapical lesions (PL) obtained with ViT models with (a) and without (b) self-supervised pretraining. (c, d) Tooth-level four-class confusion matrices (control, caries only, PL only, both) for ViT models with (c) and without (d) pretraining. (e, f) Disease-level confusion matrices for the proposed dual-head classifier (e) versus a single-head classifier (f), both built on the same pretrained ViT backbone.

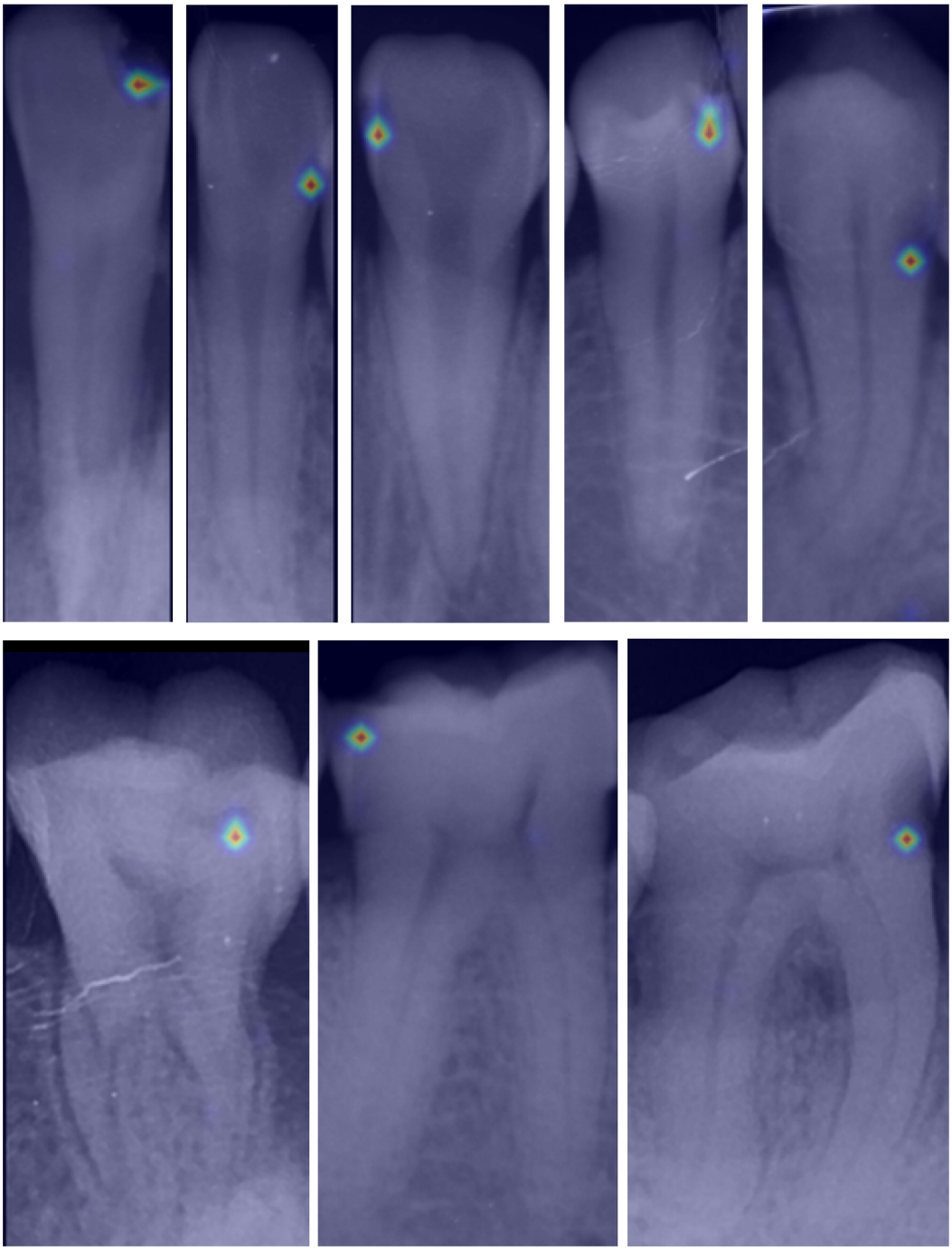

**Figure 3**

Representative attention-rollout heatmaps for **caries**on periapical radiographs obtained with the dual-head ViT model. The figure illustrates how the model identifies lesions by focusing on radiographically relevant areas in anterior teeth, premolars, and molars. The color intensity corresponds to the contribution to the prediction. High-attention areas are localized at **carious lesions.**

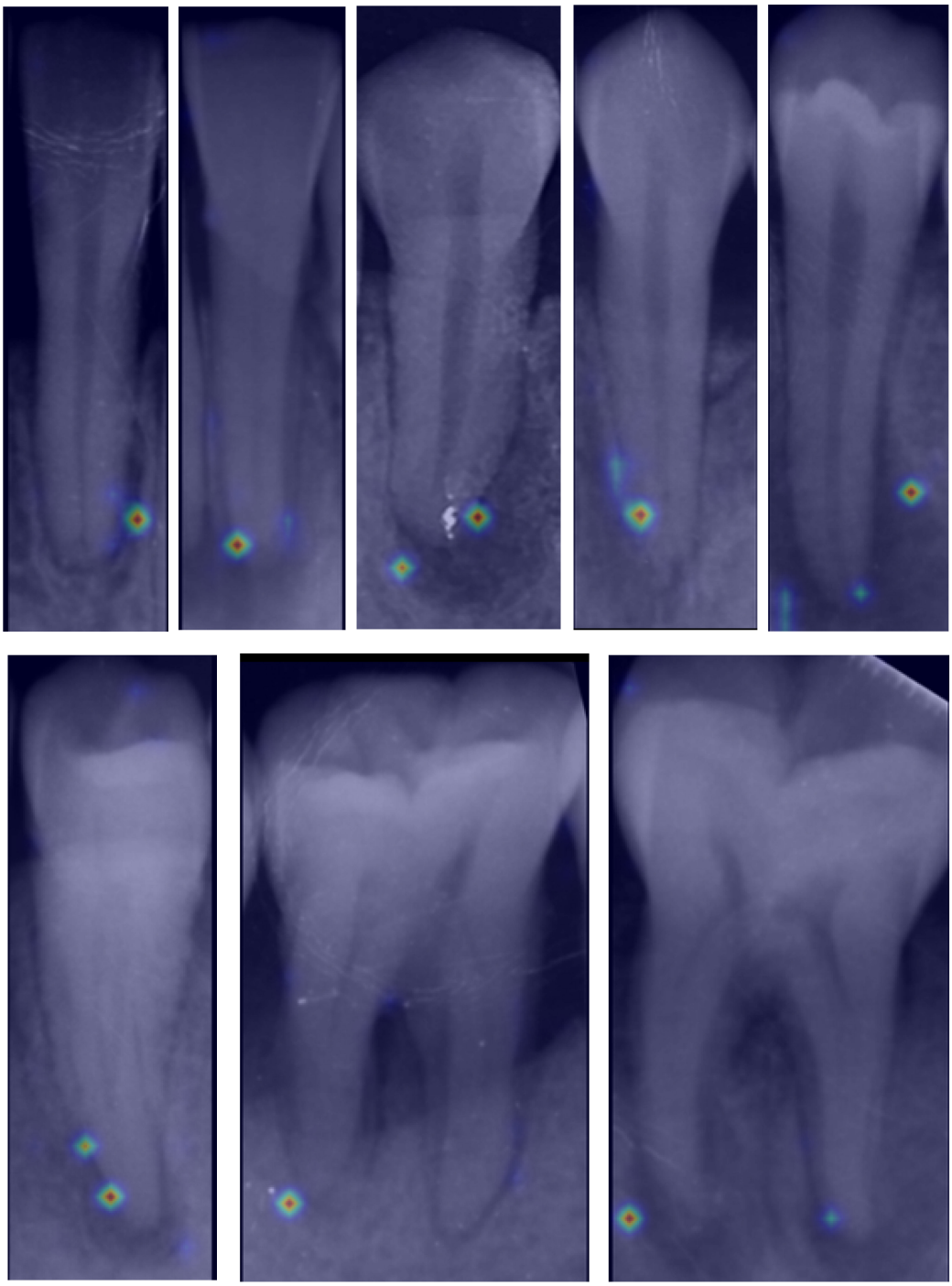

**Figure 4**

Representative attention-rollout heatmaps for periapical lesions on periapical radiographs obtained with the dual-head ViT model. The figure illustrates how the model identifies lesions by focusing on radiographically relevant areas in anterior teeth, premolars, and molars. The color intensity corresponds to the contribution to the prediction.High-attention areas are localized around the periapical regions, including multiple root apices when lesions are present in more than one root.

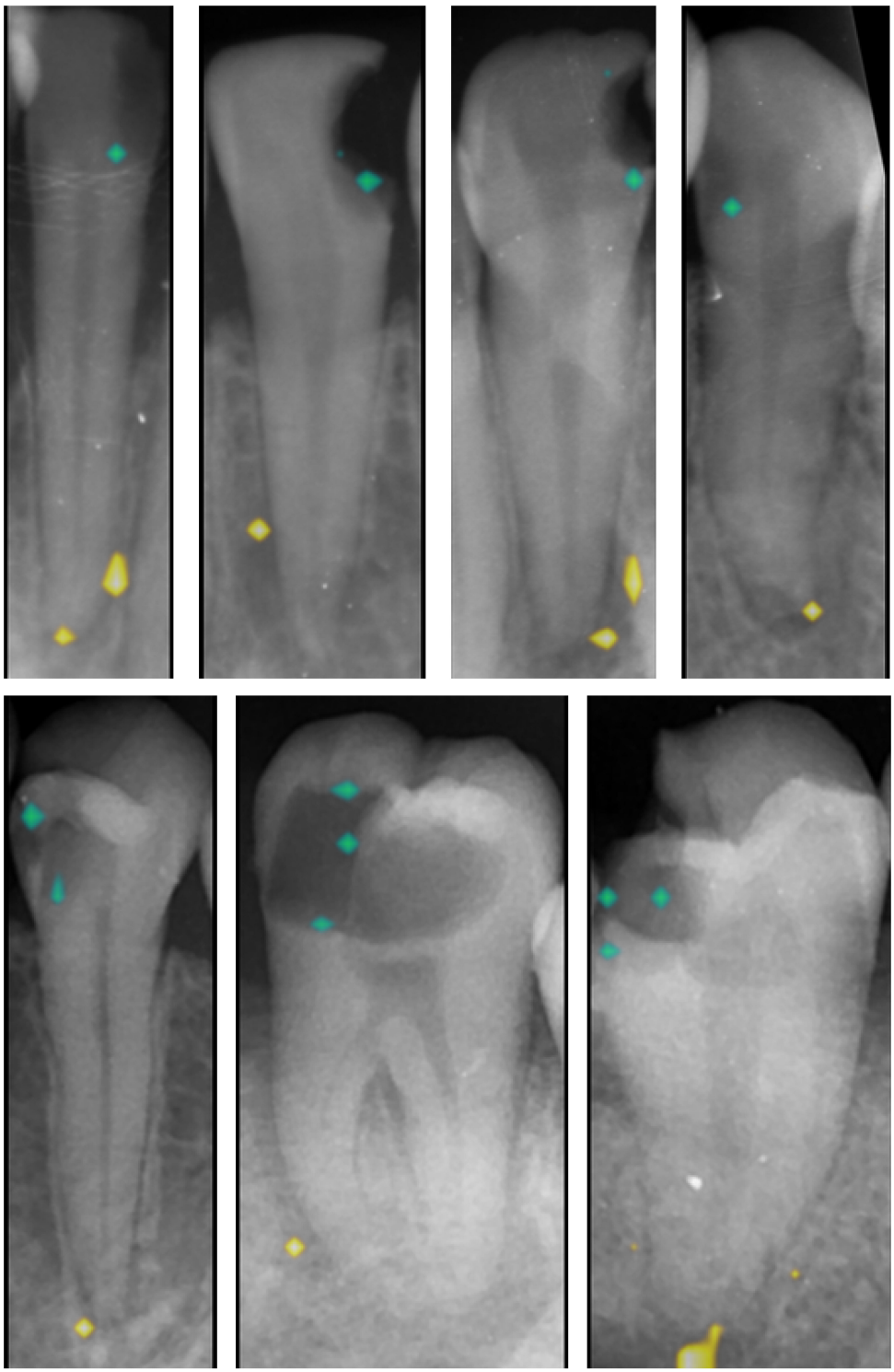

**Figure 5**

Representative attention-rollout heatmaps for **caries and** periapical lesions on periapical radiographs obtained with the dual-head ViT model. Cyan markers indicate the regions of caries, and yellow markers indicate the regions of periapical lesions, illustrating the co-existence and spatial distribution of the two conditions within the same tooth.

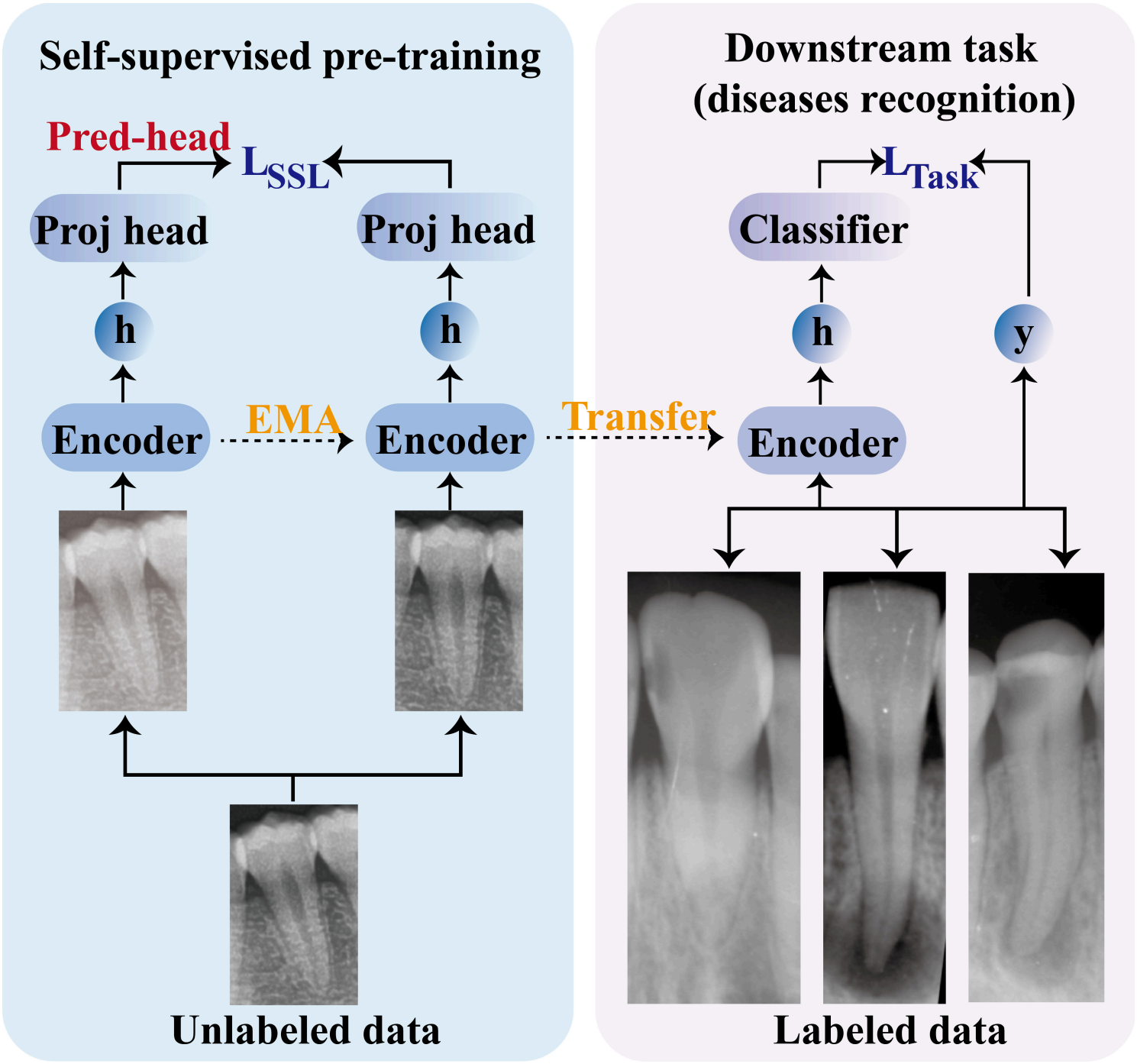


**Figure 6**

**The workflow** of the self-supervised pretraining and downstream disease-recognition task. In the self-supervised stage (left), an encoder with projection (Proj) and prediction (Pred) heads is trained on unlabeled panoramic tooth crops using an SSL loss $L_{SSL}$, and the target encoder is updated by exponential moving average (EMA). The pretrained encoder is then transferred to the downstream task (right) and fine-tuned with a classifier head on labeled periapical radiographs to predict tooth-level disease labels with the supervised loss $L_{Task}$.