ORIGINAL ARTICLE

Geometry-biased transformer for dental caries detection in panoramic X-ray images

Nima Esmi · Koosha Refahi · Asadollah Shahbahrami · Negar Khosravifard · Georgi Gaydadjiev

Received: 6 May 2025 / Accepted: 13 January 2026

Abstract
Panoramic X-ray images are essential for dental caries detection, yet they inherently suffer from geometric distortions that complicate accurate early diagnosis. To address this, we propose a Geometry-Biased transformer that explicitly models spherical geometry. Our approach integrates equirectangular relative position embedding, distance-based attention scoring, and equirectangular-aware attention rearrangement to significantly enhance spatial feature representation. This enables the model to effectively capture both local caries lesions and global dental arch structures despite distortions. We rigorously evaluated our model primarily on a hospital-scale dataset, achieving a high accuracy of 94.90% and an AUC of 0.9603. Furthermore, extensive comparative analyses across diverse publicly available dental image datasets demonstrate the superior generalization and competitive performance of our method. Our findings highlight that geometry-aware transformers offer a robust and automated tool, revolutionizing high-precision dental caries detection and potentially other medical imaging diagnostics.

Keywords Geometry-biased transformer · Swin transformer · Panoramic X-ray images · Dental caries

1 Introduction

Dental caries is one of the most prevalent chronic diseases globally, affecting approximately 24% of adults aged 20–64 years and imposing significant health and economic burdens, with annual global treatment costs exceeding billions of US dollars [1]. Early detection of caries is critical to prevent its progression from initial enamel demineralization to severe complications, such as pulp infection, abscesses, and ultimately tooth loss [2]. Panoramic X-ray imaging, or orthopantomography (OPG), serves as a cornerstone of dental diagnostics. It provides a comprehensive two-dimensional (2D) view of the entire dental arch, jaws, and surrounding structures with relatively low radiation exposure [3]. This practical and accessible modality is especially valuable for routine screenings, particularly benefiting pediatric, elderly, and disabled patients [4–7]. However, a significant challenge arises from the projection of three-dimensional (3D) dental structures onto a 2D plane. This process inherently introduces geometric distortions, such as non-linear spatial mappings, anatomical overlaps, and magnification inconsistencies, influenced by factors like patient positioning and X-ray beam divergence [8]. Consequently,

Published online: 17 March 2026

these distortions obscure subtle caries lesions, especially in their early stages, thereby challenging the precision of diagnostic assessments and highlighting the imperative for advanced computational approaches to enhance detection accuracy [9].

Traditional diagnostic methods, such as visual-tactile examinations, are limited by their inability to detect subsurface or early-stage caries and are prone to human error due to clinician fatigue and subjective interpretation [10]. To overcome these limitations, automated approaches, particularly those leveraging deep learning, have shown significant promise. Among these, Convolutional Neural Networks (CNNs) have been widely adopted for caries detection in various dental imaging modalities, excelling in local feature extraction from panoramic, bitewing, and periapical X-rays [11–13]. Nevertheless, CNNs often struggle with capturing long-range dependencies, essential for modeling the global context of panoramic radiographs, which can reduce accuracy for spatially dispersed caries [14]. Recently, transformer architectures have significantly advanced medical imaging diagnostics due to their ability to effectively model both local and global dependencies via self-attention mechanisms [15–17]. Their applications across a broad spectrum of clinical tasks including breast cancer, skin lesions, brain tumors, lung and retinal diseases, COVID-19, mental health, cardiac and colon pathologies have been systematically reviewed, consistently demonstrating superior performance over traditional convolutional neural networks in capturing long-range contextual dependencies[18–21]. Despite these advantages, standard Vision transformers, including hierarchical models like the Swin transformer, face significant challenges in robustly handling the unique spherical projection distortions inherent in panoramic X-rays. These distortions disrupt critical spatial relationships and can impair the accurate detection of subtle lesions [21]. Furthermore, recent studies highlight that these distortions lead to non-linear feature mappings, particularly at image edges, thereby reducing the robustness of transformer-based models [22]. This critically underscores the need for a geometry-aware approach capable of robustly representing spatial features despite projection-induced inconsistencies—a gap that existing methods have yet to fully address.

To address these challenges, we propose a Geometry-Biased transformer specifically tailored for dental caries detection in panoramic X-ray images. Building upon the Swin transformer architecture and leveraging principles of geometric modeling in attention mechanisms [23], our model integrates three components: equirectangular relative position embedding (ERPE), distance-based attention scoring (DAS), and equirectangular-aware attention rearrangement (EaAR). ERPE refines positional awareness by explicitly modeling spherical geometry, DAS prioritizes geometrically significant regions by adjusting attention scores based on 3D spatial relationships, and EaAR enhances global feature representation through structured information sharing across local windows. These components collectively mitigate the impact of projection distortions, leading to robust detection of both local caries lesions and global dental structures. Our model underwent comprehensive evaluations, primarily on the hospital-scale CariesXrays dataset [24], where it achieved a high accuracy of 94.90% and an AUC of 0.9603. Furthermore, comparative assessments on other publicly available dental imaging datasets demonstrate the broader applicability and competitive performance of our proposed method, notably outperforming existing CNNs and baseline Swin transformers. This significant improvement highlights the potential of geometry-aware transformers to advance automated dental diagnostics. The remainder of this article is structured as follows: Sect. 2 reviews background and related work. Section 3 details the proposed model architecture and evaluation datasets. Sections 4 and 5 present evaluation metrics, results, and an ablation study. Section 6 discusses findings.

2 Background and related work

This section provides an overview of fundamental imaging modalities used in dental diagnostics and summarizes key developments in transformer-based methods for medical and dental image analysis.

2.1 Background information

2.1.1 Types of dental images used in AI analysis

A variety of dental radiographic modalities are employed in AI-driven diagnostic systems. The most commonly used types, illustrated in Fig. 1, include bitewing, periapical, occlusal, cephalometric, cone-beam computed tomography (CBCT), and panoramic radiographs [25–29]. Each modality provides complementary clinical information relevant to different diagnostic tasks.

- **Bitewing:** Captures the crowns of posterior teeth and adjacent bone levels, primarily used for detecting proximal caries and assessing periodontal status.
- **Periapical:** Provides a complete view of the tooth from crown to root apex, supporting diagnosis of periapical lesions, root fractures, and endodontic complications.
- **Occlusal:** Offers a wider view of the occlusal plane, useful for detecting developmental anomalies, supernumerary teeth, and jaw expansion patterns.
- **Cephalometric:** Delivers lateral or frontal skull projections, widely used in orthodontics to evaluate craniofacial morphology and growth patterns.
- **CBCT:** Produces three-dimensional volumetric data enabling detailed visualization of bone morphology, implant planning, and evaluation of complex anatomical structures.
- **Panoramic:** Provides a holistic two-dimensional representation of the entire oral cavity, including both dental arches, temporomandibular joints, and surrounding structures.

Among these modalities, panoramic radiographs hold particular importance for automated dental analysis. They are routinely acquired in clinical workflows, involve relatively low radiation compared to CBCT, and capture a comprehensive view that allows simultaneous assessment of caries, alveolar bone levels, impacted teeth, and jaw pathologies. Their broad coverage and widespread availability make them highly suitable for large-scale AI-driven diagnostic systems.

2.1.2 Transformers and image analysis

Transformer architectures were originally introduced for natural language processing, where self-attention enables efficient modeling of long-range dependencies [30]. Their success led to adaptations in computer vision, most notably the Vision Transformer (ViT), which treats fixed-size image patches as tokens and applies global self-attention to learn both local and contextual features [16].

In dental imaging, ViTs offer advantages over convolutional neural networks (CNNs) by capturing global spatial relationships across panoramic radiographs–an essential property for detecting subtle or spatially distributed caries lesions. Their robustness to variations in image resolution and anatomical overlap has positioned ViTs as a strong alternative to traditional CNN-based systems.

The Swin Transformer further improved computational efficiency by introducing hierarchical representations and window-based self-attention [31]. This design significantly reduces computational cost, making it practical

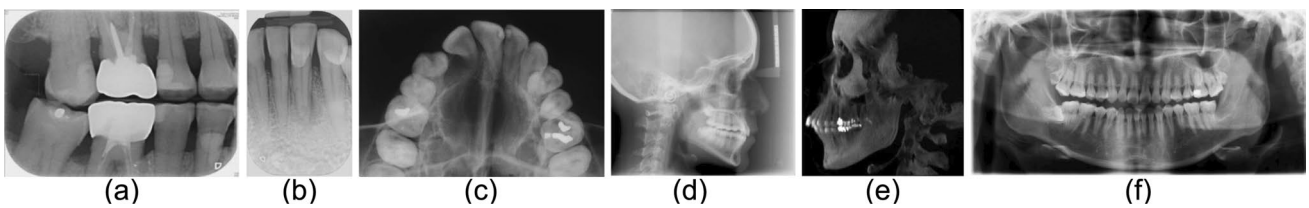

Fig. 1 Common types of dental images used in AI analysis: **a** Bitewing, **b** Periapical, **c** Occlusal, **d** Cephalometric, **e** Cone Beam Computed Tomography (CBCT), and **f** Panoramic

for high-resolution medical images, including panoramic X-rays. Swin-based models have demonstrated strong performance across segmentation, classification, and detection tasks in medical imaging.

Despite these strengths, transformer models encounter challenges when applied to panoramic radiographs. The projection of three-dimensional dental anatomy onto a two-dimensional plane introduces geometric distortions, especially toward the periphery of the image. These distortions disrupt spatial consistency and degrade the ability of standard self-attention mechanisms to accurately model anatomical relationships. As highlighted in recent studies [22, 23], these nonlinear projection effects can obscure early caries lesions and reduce detection accuracy.

Consequently, transformer architectures for panoramic X-ray analysis benefit from geometry-aware adaptations that explicitly account for the underlying spherical imaging geometry. Such adaptations motivate the development of specialized models, including the geometry-biased transformer proposed in this work.

2.2 Related work

Panoramic X-ray imaging is a cornerstone of modern dentistry, offering a comprehensive two-dimensional view of the oral and maxillofacial region, including teeth, jaws, and surrounding anatomical structures. This modality is extensively used for diagnosing dental caries, periodontal diseases, impacted teeth, and other pathologies due to its broad coverage [8, 32, 33]. However, manual interpretation of panoramic radiographs is labor-intensive, subjective, and susceptible to errors, particularly in identifying early-stage caries or subtle lesions. Deep learning has emerged as a transformative approach, automating diagnostic tasks with high accuracy and efficiency. Studies such as [34] and [35] underscore the growing interest in deep learning for dental caries detection, highlighting its potential to enhance clinical decision-making, reduce diagnostic variability, and meet the demands of modern dental healthcare.

A significant body of research has leveraged CNNs for detecting dental caries and related conditions in panoramic and periapical X-rays. Szabo et al. [34] proposed a deep learning pipeline that combines YOLOv3 for lesion localization and MobileNetV2 for classification, achieving promising results in detecting periapical lesions on panoramic radiographs. Similarly, Singh et al. [35] developed an automatic framework based on Faster R-CNN with a ResNet-50 backbone, which effectively localizes carious lesions with high accuracy. Anbarasi et al. [36] improved CNN-based detection by incorporating preprocessing techniques such as histogram equalization and contrast adjustment, enhancing caries identification. Schneider et al. [21] evaluated the performance of CNNs and reported a precision of 0.83 ± 0.22 for caries segmentation, outperforming transformer and hybrid models in data-limited scenarios. Ying et al. [37] compared models like YOLOv5 and UNet against dentists, finding that YOLOv5 achieved a Youden index of 0.76, closely rivaling human performance. Ayhan et al. [38] utilized YOLOv7 and an enhanced YOLOv7+CBAM model for caries detection under fixed dental prostheses, achieving a recall of 0.827 and a mean average precision (mAP) of 0.846. These CNN-based approaches demonstrate excellence in local feature extraction and robust performance across various dental imaging tasks.

Despite their successes, CNN-based models exhibit certain limitations when applied to panoramic X-rays. The large spatial dimensions of these images often necessitate aggressive downsampling, which can lead to the loss of fine details crucial for early caries detection [14]. Moreover, CNNs primarily focus on local feature extraction and lack the ability to model long-range dependencies, which are essential for capturing the global context of panoramic radiographs. In contrast, ViTs with their self-attention mechanisms, overcome these limitations by effectively modeling spatial relationships across the entire image. For instance, studies by Van et al. and Hao et al. [22, 39] demonstrate that ViTs excel in processing large-scale images and capturing holistic dependencies, providing a significant advantage over CNNs for complex dental imaging tasks. This shift towards ViTs marks a promising evolution in automated dental diagnostics. transformer-based models have gained traction in dental imaging, often combined with CNNs or enhanced with novel strategies. Van et al. [39] proposed a hybrid CNN-transformer model for caries detection, integrating local and global feature extraction to achieve improved performance. Sun et al. [40] enhanced the TransUNet architecture by incorporating attention mechanisms and self-training, resulting in superior tooth segmentation on the MICCAI STS-2D dataset. Schneider et al. [21]

compared transformers with CNNs and hybrid models, highlighting their potential despite their data-intensive requirements. Ying et al. included DEtection TRansformer (DETR) [41] and Trans-UNet in their comparison, reporting mixed results when pitted against CNNs [37]. Kuccuk et al. [42] developed a hybrid CNN-transformer model utilizing YOLOv8 and RT-DETR for impacted tooth detection. Alsakar et al. [43] combined MobileNetV2 with a transformer transformer and a bagging ensemble. Kim et al. [44] employed a convolutional variational autoencoder and clustering for periodontal disease diagnosis, outperforming supervised vision transformers by up to 14%. Hao et al. [22] introduced SemiTNet, a semi-supervised transformer framework, achieving an Intersection over Union (IoU) of 94.41% and a Dice score of 95.45% for tooth segmentation on the TSI15k dataset. These works underscore the versatility and power of transformers in advancing dental diagnostics.

Transformer-based models, despite their strengths, encounter challenges in panoramic X-rays due to distortions arising from projecting a three-dimensional spherical anatomy onto a two-dimensional plane. These distortions disrupt spatial feature consistency, as observed by [14], impacting the performance of ViTs in detecting subtle caries or lesions obscured by anatomical warping [14]. Schneider et al. and Hao et al. [21, 22] acknowledge that transformers struggle with such inconsistencies, particularly in data-scarce scenarios or when distortions are pronounced. Although hybrid approaches, such as those proposed by Van et al. and Alsakar et al. [39, 43] mitigate some of these issues by blending CNN and transformer strengths, the challenge remains unresolved.

Addressing these projection-related challenges necessitates tailored modifications to transformer architectures to ensure robust feature representation despite panoramic distortions. In this regard, significant advancements have been made in related fields concerning the processing of distorted spherical (e.g., 360-degree) images, which share similar geometric challenges with panoramic X-rays. For instance, Yun et al. [23] introduced EGformer, an equirectangular geometry-biased transformer specifically designed for 360-degree depth estimation, which leverages equirectangular-aware attention mechanisms to counteract projection distortions. Similarly, for 360-degree image classification, Cho et al. [45] proposed a sampling-based spherical transformer that directly samples pixels from the sphere surface, thereby circumventing erroneous planar projection processes and enhancing model performance. These works underscore the broader need for specialized geometry-aware transformer designs to effectively handle the unique spatial properties of distorted input images.

3 Proposed approach

The proposed approach, including the architecture of the Geometry-Biased transformer and its evaluation datasets, are presented in this section.

3.1 Model architecture

A representative dental Panoramic X-ray image is depicted in Fig. 2c. Such images often require adjusting their dimensions to project a quasi-spherical space Fig. 2a onto a plane Fig. 2b. This transformation alters the distances between points across different parts of the image, introducing spatial distortions. The transformer transformer was designed to process Panoramic X-ray images effectively and make all the changes [34]. Vision transformer models typically partition images into uniform dimensions to generate image tokens Fig. 3a. However, when applied to dental Panoramic X-ray images (as illustrated in Fig. 2), the spatial relationships between image tokens deviate from linearity Fig. 3b. This non-linear mapping can mislead the model, impacting its performance and accuracy. To address this issue, inspired by [46, 47], and [23], we have introduced the architecture depicted in Fig. 4, which consists of three main components: the encoder, bottleneck, and decoder. Additionally, the details of the transformer block are illustrated at the bottom of the figure, and these will be elaborated upon in the following sections. The model architecture combines a symmetric encoder-decoder design with skip connections to effectively learn and reconstruct spatial and contextual features. The encoder splits input images into non-overlapping patches, treating each patch as a token. These tokens pass through transformer transformer blocks, which

 https://doi.org/10.1007/s00521-026-11939-x

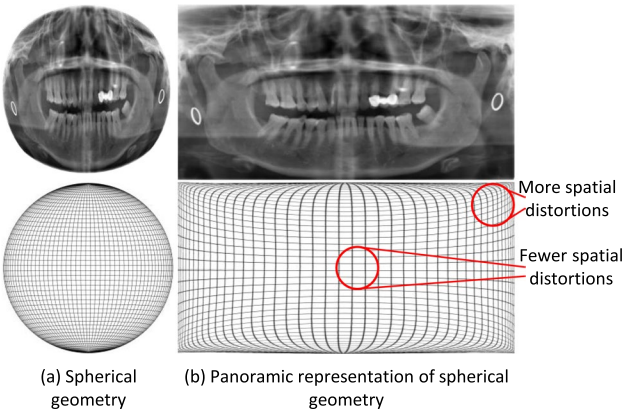

Fig. 2 Formation of dental Panoramic X-ray images through the flattening of spherical geometry images. **a** The concept of spherical geometry, **b** Representing spherical concepts on a plane

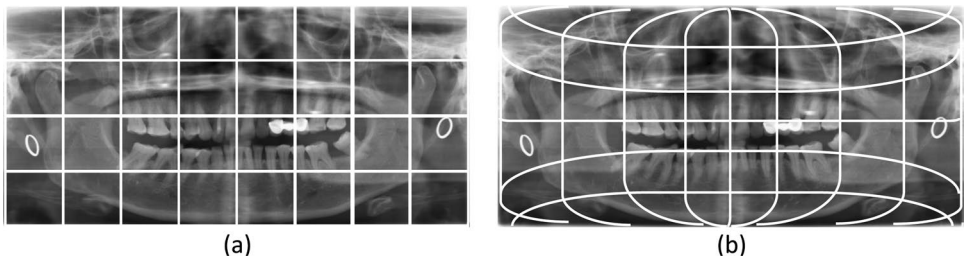

Fig. 3 **a** Partitioning a vision transformer into equally-sized sections for tokenizing images. **b** Non-linear dependency of tokens in Panoramic images

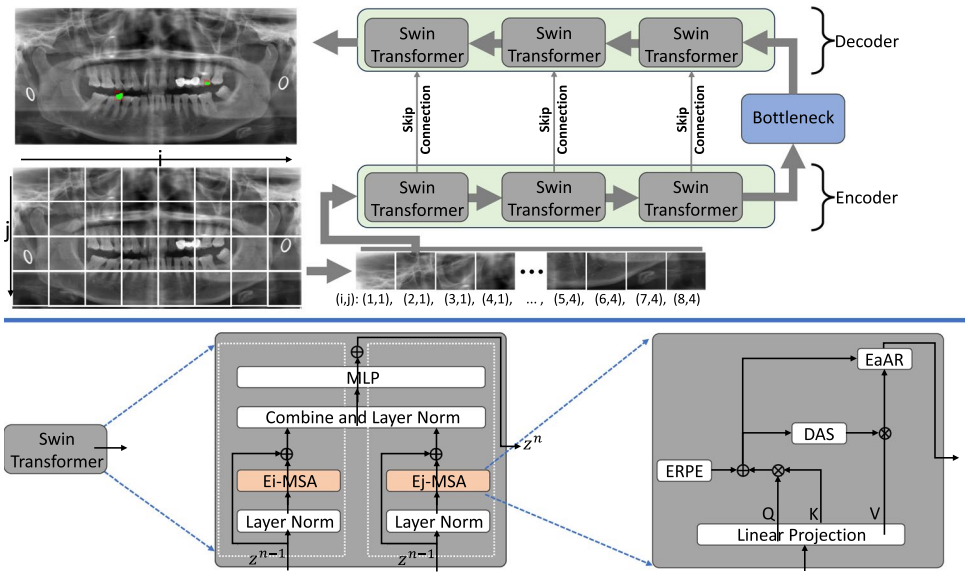

Fig. 4 The Geometry-Biased transformer model architecture consists of three main components: Encoder, Bottleneck, and Decoder. Each transformer block includes two attention mechanisms: horizontal (Ei-MSA) and vertical (Ej-MSA), controlled by the mechanisms DAS, ERPE, and EaAR

utilize shifted window-based self-attention to capture both local and global semantic information. Down-sampling can be done while increasing feature dimensions, creating a hierarchical representation. At the bottleneck, two consecutive transformer transformer blocks further refine deep feature representations. The decoder mirrors the encoder, employing a patch expanding layer to up-sample features and reduce dimensions, restoring spatial resolution. Skip connections fuse multi-scale features from the encoder with the decoder's up-sampled outputs, mitigating information loss from down-sampling. This architecture eliminates convolutional operations, relying entirely on transformers for both feature extraction and spatial reconstruction, making it capable of modeling long-range dependencies crucial for precise segmentation. Swin-Unet's design enhances performance in complex tasks by achieving superior accuracy and boundary delineation.

3.2 Horizontal and vertical attitude

In a transformer transformer block, attention is computed locally within non-overlapping windows, which are processed sequentially in horizontal and vertical directions. The horizontal part focuses on attending to patches of the input image that are grouped along the horizontal (i) axis. It processes these windows by computing the self-attention among the patches within each window, allowing for a fine-grained understanding of spatial relationships along the horizontal dimension. Similarly, the vertical part handles vertical slices of the image, computing self-attention along the vertical (j) axis. This complementary structure, where both horizontal and vertical directions are attended to separately, helps the model capture interactions along both axes of the image. To enhance this, the windows shift between layers (shifted window mechanism), ensuring that the windows in one layer overlap with the windows in the next, allowing the model to aggregate information globally over multiple layers despite the local nature of attention within each block. This structure leads to efficient computation with fewer parameters while maintaining strong spatial representation capabilities across both axes. Equirectangular i axis Multi-Head Self-Attention(Ei-MSA) and Equirectangular j axis Multi-Head Self-Attention (Ej-MSA) are core components of the geometry-biased transformer. They work together to process equirectangular images by applying attention mechanisms separately along the horizontal and vertical axes, accounting for the unique geometric distortions in 360-degree images. They consist of three main parts: ERPE, DAS, and EaAR.

3.2.1 Equirectangular relative position embedding

The ERPE introduces a bias that accounts for the unique geometry of equirectangular images, which are projections of spherical images. In these projections, spatial relationships are not uniform, meaning that the distance between pixels (or image elements) in 2D space does not necessarily reflect their actual 3D spatial relationships. ERPE measures the relative position between pixels in a local window using spherical coordinates. For each local window, the attention score is modified based on these relative positions in 3D space. The embedding adjusts attention scores to give higher weight to pixels that are closer together in 3D space, even if they appear further apart in 2D. For example, pixels on the left and right edges of the equirectangular image are actually next to each other in the spherical space. This helps to improve transformer ability to handle the distortions introduced by the equirectangular projection.

3.2.2 Distance-based attention score

DAS replaces the traditional softmax function used in attention mechanisms. Softmax is computationally expensive and does not naturally reflect the geometric distances in equirectangular images. Instead of just calculating attention based on pixel values, it takes into account the actual 3D distance between pixels. The further an element is from a baseline point in the spherical space, the higher the attention score it receives. This baseline is carefully chosen to ensure balanced attention across the image. This process amplifies the differences in scores, ensuring that important regions (closer to the baseline) receive more focus, while regions that are far from the baseline (geometrically) receive less attention. The detailed explanation of DAS is presented in Algorithm 1. The algorithm starts by constructing query (Q), key (K), and value (V) matrices, along with spherical parameters that define the geometric relationships in the data. Using ERPE, it modifies traditional attention scoring to embed positional biases based on spherical geometry. The attention score is computed by combining the dot product of Q and K with the ERPE, yielding an angle-based metric ($\Delta\theta$) normalized to spherical coordinates. This metric is used to calculate a geometric distance (d_{geo}) that reflects spatial relations on a sphere. DAS leverages d_{geo} as the attention weight, ensuring that the resulting attention scores prioritize regions with significant geometric relationships. Each pixel's attention is adjusted by weighting its corresponding value (V) matrix entry with d_{geo} . This approach ensures that the attention mechanism accounts for the inherent geometric distortions of spherical

projections, making it particularly effective for tasks like 360-degree image analysis or depth estimation. The algorithm's final output is a refined attention map that integrates both feature similarity and geometric context.

Algorithm 1 Distance-based attention score Algorithm

Require:
$Q \in \mathbb{R}^{1\times W\times d_j}$ ▷ Query matrix of size (1, width, hidden dimension)
$K \in \mathbb{R}^{1\times W\times d_j}$ ▷ Key matrix of size (1, width, hidden dimension)
$V \in \mathbb{R}^{1\times W\times d_j}$ ▷ Value matrix of size (1, width, hidden dimension)
$\phi \in \mathbb{R}^1$ ▷ Angle parameter representing the vertical position in spherical coordinates
ρ_b, θ_b, ϕ_b ▷ Baseline spherical coordinates parameters
$\mathrm{ERPE}(Q,K) \in \mathbb{R}^{W\times W}$ ▷ Equirectangular Relative Position Embedding
Ensure: Distance-based attention score matrix $\mathrm{DAS} \in \mathbb{R}^{W\times W}$
procedure ComputeDAS
  $\mathrm{score}_h \leftarrow QK^T + \mathrm{ERPE}(Q,K)$ ▷ Calculate the attention score with ERPE
  $\Delta\theta \leftarrow N(\mathrm{score}_h)\cdot\frac{\pi}{2}$ ▷ Normalize and scale the score to spherical angles
  $d_{\mathrm{geo}} \leftarrow 2\rho_b^2\cdot(1-\cos(\Delta\theta))\cdot\sin^2(\phi_b)$ ▷ Geometric distance in spherical coordinates
  $\mathrm{DAS} \leftarrow d_{\mathrm{geo}}$ ▷ Use the geometric distance as the attention score
  for all pixels i in the local window **do**
    $\mathrm{Attention}_h[i] \leftarrow \mathrm{DAS}[i]\cdot V[i]$ ▷ Weight the value matrix by DAS
  end for
  return $\mathrm{Attention}_h$ ▷ Return the final attention scores
end procedure

3.2.3 Equirectangular-aware attention rearrangement

While the ERPE and DAS components improve local attention, they do not allow for direct interaction between different local windows (sets of pixels being attended to). EaAR addresses this by enabling local windows to interact with one another indirectly. Each local window (either horizontal or vertical) is assigned an importance score based on its geometric characteristics (using DAS and ERPE). EaAR rearranges the attention scores of these local windows, allowing information to flow between windows. This creates an effect similar to having a larger, global receptive field (even though the model still uses local attention). Windows that are deemed important get a higher weight in the attention process, while less important windows receive less attention. EaAR ensures that the final attention score for each local window is influenced not only by the local context but also by the broader image structure. It combines attention from multiple windows and incorporates this information into the final output.

3.2.4 Computational complexity analysis

The design of our geometry-biased transformer carefully considers computational efficiency, particularly within its attention mechanisms. While standard global attention mechanisms, with a quadratic complexity of $O((HW)^2)$ with respect to the input resolution, are computationally prohibitive for high-resolution images, our architecture leverages principles inspired by efficient transformers for handling distorted inputs, such as those in equirectangular projection [23]. Instead of a full global attention, our model incorporates a form of equirectangular-aware local attention. Specifically, our attention mechanism operates by modeling interactions along specific axes to account for the unique geometric distortions in panoramic X-ray images. This approach avoids the full quadratic complexity of global attention by focusing computations. The computational complexity of our attention mechanisms, namely the Ei-MSA and Ej-MSA, can be formally expressed as:

$$\Omega(\text{Ei-MSA}) = 4HWC^2 + 2HW^2C$$
$$\Omega(\text{Ej-MSA}) = 4HWC^2 + 2H^2WC$$

where H is the height, W is the width, and C is the channel dimension. These formulas highlight that while a term quadratic with respect to one spatial dimension (H^2 or W^2) is present – which is necessary to capture global dependencies along an axis for accurate geometric correction – the overall complexity remains significantly lower than that of global attention. Our proposed model has approximately 15.5 million parameters and a computational complexity of approximately 74 *GFLOPs*. Furthermore, it demonstrates an average inference time of

400 ms per panoramic X-ray image when executed on a system equipped with an Intel Core i7 CPU, 16 *GB* of RAM, and an NVIDIA GeForce GTX 1080 Ti GPU. This design ensures accurate modeling of panoramic distortions while maintaining a more favorable balance between computational burden and the achieved diagnostic performance, making it suitable for clinical deployment.

4 Evaluation metrics and results

In this section, performance evaluation criteria, some datasets, evaluation results and ablation study are presented.

4.1 Performance evaluation criteria

The proposed method was evaluated using six key metrics: accuracy, precision, recall, F1-score, False Negative Rate (FNR), and area-under-the-curve (AUC), which are commonly employed in classification tasks. These metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{FNR} = \frac{FN}{TP + FN} \tag{5}$$

The False Negative Rate (FNR) quantifies the proportion of actual positive cases that are incorrectly classified as negative, representing missed detections.

The AUC quantifies the model's ability to distinguish between classes by calculating the area under the receiver operating characteristic (ROC) curve. To benchmark performance, a comparative analysis with two attending dentists was conducted, including kappa coefficients to assess inter-observer agreement. The kappa coefficient, computed as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{6}$$

measures agreement, where P_o is the observed agreement and P_e is the expected agreement by chance. Kappa values ranging from 0.6 to 1.0 indicate substantial to nearly perfect agreement between the dentists and the model.

Additionally, the time required for diagnosis by the proposed model was significantly lower than manual evaluations, highlighting its efficiency. Overall, the evaluation demonstrated that the model achieves high classification performance, particularly in caries diagnosis for primary molars, underscoring its potential as a reliable tool for clinical use.

4.2 Datasets

The **CariesXrays** dataset [24] was used as the primary benchmark. It contains 6000 hospital-collected panoramic dental X-ray images (resolution: 1333×800 pixels) with 13,783 expert-annotated caries instances from three dental professionals, reflecting real clinical variability. Evaluation was performed using 5-fold cross-validation.

The **MICCAI STS-2D Challenge** dataset (MICCAI 2024) [48] provides expertly annotated 2D panoramic X-ray images for tooth segmentation, accompanied by a larger set of unlabeled images to support semi-supervised learning.

The **Dental Radiography** dataset [49] consists of 851 panoramic dental X-ray images with expert annotations for caries and other dental abnormalities.

The **Panoramic Dental Dataset** [50] includes 116 high-resolution panoramic X-ray images (≈2000×1000 pixels) with pixel-level caries annotations by dental experts.

The **TSI15k** dataset [51] comprises 15,000 panoramic dental X-ray images (typically 1024×512 pixels), of which a subset is expertly labeled for tooth segmentation and identification.

The **DENTEX** dataset (MICCAI 2023) [52] contains 5000 panoramic dental X-ray images (≈1500×700 pixels) annotated with tooth positions, numbers, and diagnostic labels including caries.

4.3 Evaluation results

To evaluate the performance of the geometry-biased transformer for dental caries detection, we compared it against four widely used CNNs (AlexNet, GoogleNet, SeNet, and ResNet), the traditional transformer transformer, and

Table 1 Performance comparison between our proposed approach (Geometry-biased transformer), different CNN baselines and baseline transformer transformers on CariesXrays, MICCAI STS-2D, Dental Radiography, Panoramic Dental, TSI15k, and DENTEX datasets

Methods	Accu.%	Pre.%	Rec.%	F1.%	FNR.%	AUC
CariesXrays						
AlexNet	63.30	65.56	64.78	65.03	35.22	0.6819
GoogleNet	67.23	66.67	70.14	71.67	29.86	0.6978
SeNet	76.23	78.71	75.13	78.09	24.87	0.8001
ResNet	78.26	79.65	80.14	81.55	19.86	0.8312
[21], DeepLabV3+	80.13	81.05	81.94	82.09	18.06	0.8407
[37], DETR	81.29	82.11	83.03	83.89	16.97	0.8594
Swin transformer	84.57	86.09	84.16	84.56	15.84	0.8778
[45]	90.83	91.50	90.50	91.00	9.50	0.9200
Ours	**94.90**	**97.60**	**95.52**	**94.84**	**4.48**	**0.9603**
MICCAI STS-2D						
[40], STS-TransUNet	70.17	70.01	68.55	69.32	31.45	0.7335
[44], UNet	72.38	71.70	73.61	73.81	26.39	0.7724
[45]	80.02	81.00	80.50	80.75	19.50	0.8200
Ours	**83.90**	**87.00**	**86.33**	**85.17**	**13.67**	**0.8652**
Dental Radiography						
[43], MobileNetV2	94.08	93.27	93.14	93.17	6.86	0.9338
[45]	95.55	95.00	94.50	94.75	5.50	0.9450
Ours	**97.77**	**97.61**	**94.54**	**96.90**	**5.46**	**0.9553**
Panoramic Dental Dataset						
[36], DentSU_Net	97.01	95.83	95.05	95.66	4.95	0.9407
[45]	96.17	96.00	95.50	95.75	4.50	0.9350
Ours	**97.80**	**97.00**	**96.19**	**96.64**	**3.81**	**0.9503**
TSI15k						
[22], SemiTNet	≈92.00	94.74	**97.10**	95.90	2.90	–
[45]	93.57	94.00	94.50	94.25	5.50	0.9300
Ours	**96.11**	**95.90**	**96.05**	**96.00**	**3.95**	**0.9477**
DENTEX						
[39]	≈94.70	–	–	71.20	–	–
[45]	94.90	94.50	94.00	94.25	6.00	0.9350
Ours	**96.11**	**95.90**	**96.05**	**96.00**	**3.95**	**0.9477**

hybrid models. As shown in Table 1, the geometry-biased transformer significantly outperformed all geometry-unbiased baselines, demonstrating the potential of this approach for improving the accuracy of caries detection on panoramic X-ray images. Bold values represent the highest score in each column.

To rigorously evaluate the generalizability and robust performance of our proposed model across diverse dental imaging contexts, we conducted a comprehensive 5-fold cross-validation on six distinct datasets: CariesXrays, MICCAI STS-2D, Dental Radiography, Panoramic Dental Dataset, TSI15k, and DENTEX. The detailed results, including Accuracy, Precision, Recall, and F1-Score for each fold and their overall averages per dataset, are presented in Table 2. As demonstrated by the results, our model consistently maintains high performance across the majority of these benchmarks. Specifically, on the CariesXrays dataset, the model achieved an average (in bold) Accuracy of 94.90% and an F1-Score of 94.84%. Similarly, high average performance was observed on the

Table 2 5-Fold Cross-Validation Performance of Our Model Across Diverse Dental Image Datasets (Mean ± Standard Deviation)

Methods	Accu.%	Pre.%	Rec.%	F1.%
CariesXrays				
Fold 1	94.75	97.55	95.40	94.70
Fold 2	95.10	97.70	95.60	94.95
Fold 3	94.80	97.50	95.50	94.80
Fold 4	95.00	97.65	95.55	94.90
Fold 5	94.85	97.60	95.55	94.85
Average	**94.90**	**97.60**	**95.52**	**94.84**
MICCAI STS-2D				
Fold 1	83.70	86.80	86.20	85.00
Fold 2	84.10	87.20	86.40	85.30
Fold 3	83.85	86.95	86.30	85.15
Fold 4	83.95	87.05	86.45	85.25
Fold 5	83.90	87.00	86.30	85.15
Average	**83.90**	**87.00**	**86.33**	**85.17**
Dental Radiography				
Fold 1	97.70	97.50	94.40	96.80
Fold 2	97.80	97.70	94.60	97.00
Fold 3	97.75	97.60	94.50	96.85
Fold 4	97.85	97.65	94.65	96.95
Fold 5	97.75	97.60	94.55	96.90
Average	**97.77**	**97.61**	**94.54**	**96.90**
Panoramic Dental Dataset				
Fold 1	97.70	96.90	96.10	96.50
Fold 2	97.90	97.10	96.25	96.70
Fold 3	97.75	96.95	96.15	96.60
Fold 4	97.85	97.05	96.20	96.75
Fold 5	97.80	97.00	96.25	96.65
Average	**97.80**	**97.00**	**96.19**	**96.64**
TSI15k				
Fold 1	96.00	95.80	95.90	95.90
Fold 2	96.20	96.00	96.15	96.10
Fold 3	96.10	95.85	96.00	95.95
Fold 4	96.15	95.95	96.10	96.05
Fold 5	96.10	95.90	96.10	96.00
Average	**96.11**	**95.90**	**96.05**	**96.00**
DENTEX				
Fold 1	96.05	95.85	96.00	95.95
Fold 2	96.15	95.95	96.10	96.05
Fold 3	96.10	95.90	96.05	96.00
Fold 4	96.00	95.80	95.95	95.90
Fold 5	96.35	96.00	96.15	96.10
Average	**96.11**	**95.90**	**96.05**	**96.00**

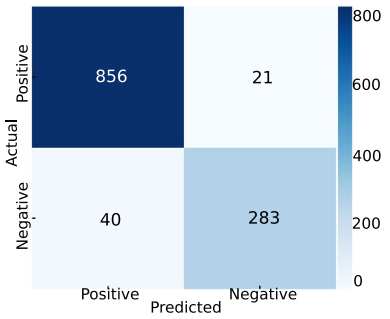

Fig. 5 Confusion matrix based on 1200 test samples on CariesXrays dataset, illustrating the model's performance in detecting dental caries

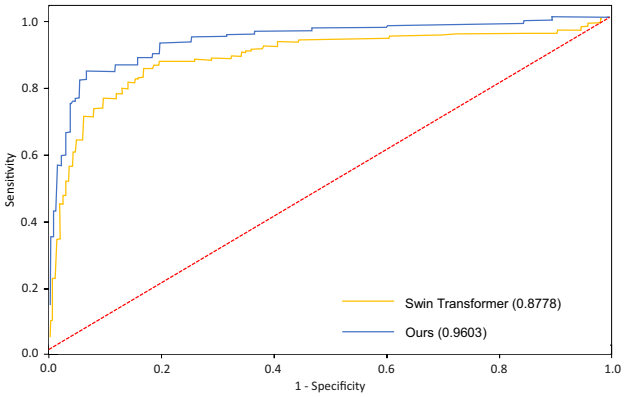

Fig. 6 Receiver operating characteristic curves of the proposed geometry-biased transformer and the baseline swin transformer, on CariesXrays dataset. Numbers in parentheses show the AUC values of the two methods

Dental Radiography (Acc: 97.77%, F1: 96.90%), Panoramic Dental Dataset (Acc: 97.80%, F1: 96.64%), TSI15k (Acc: 96.11%, F1: 96.00%), and DENTEX (Acc: 96.11%, F1: 96.00%) datasets. While the MICCAI STS-2D dataset showed comparatively lower average performance (Acc: 83.90%, F1: 85.17%), the consistent and high scores across the other five diverse datasets strongly affirm the model's remarkable generalization ability and its reliability for practical application in varied clinical scenarios, despite inherent variations in image characteristics and acquisition protocols.

Additionally, Fig. 5 shows the confusion matrix for our approach on CariesXrays dataset. These results highlight the model's robust performance in correctly identifying dental caries (TP) while minimizing false positives (FP) and false negatives (FN). The high precision value (0.976) reflects the model's ability to accurately classify positive cases, ensuring minimal overdiagnosis. Similarly, the recall (0.9552) underscores its effectiveness in detecting actual cases of dental caries, reducing the risk of underdiagnosis.

By incorporating spatial relationships between image tokens, the proposed geometry-biased transformer achieves approximately a 10% accuracy improvement over the baseline transformer transformer. Figure 6 illustrates the ROC curves of our model and baseline transformer transformer, further demonstrating the efficacy of incorporating domain knowledge into caries detection.

5 Ablation study

To systematically evaluate the individual contribution of each proposed component to the overall performance of our geometry-biased Transformer, we conducted an ablation study on the CariesXrays dataset. Table 3 presents the accuracy performance of various model configurations over 10 repeated runs, along with their mean, standard deviation (Both are marked as bold numbers), and *p*-values (indicating statistical significance against the

Table 3 Accuracy performance (in percentage) of model variations in an ablation study on the CariesXrays dataset, evaluated over 10 repeated runs

Run/Metric	B	B + ERPE	B + ERPE + DAS	B + ERPE + DAS + EaAR
Run 1	84.14	90.82	94.75	94.75
Run 2	84.61	89.97	93.95	95.10
Run 3	84.58	90.89	93.33	94.80
Run 4	84.07	90.25	93.15	95.00
Run 5	87.45	87.42	90.89	94.63
Run 6	85.07	90.05	94.90	95.10
Run 7	84.90	91.50	93.65	94.75
Run 8	83.88	92.08	94.88	94.95
Run 9	83.40	91.22	94.65	94.67
Run 10	84.50	90.00	93.25	94.05
Mean	**84.56**	**90.42**	**93.74**	**94.79**
Std. Dev.	**0.48**	**1.28**	**1.26**	**0.31**
p-value (vs. B)	N/A	<0.001	<0.001	<0.001

The table displays individual accuracy scores per run, along with the mean, standard deviation, and *p*-value (vs. Baseline B, using Wilcoxon Signed-Rank test). B, Baseline transformer-Unet transformer; ERPE, Equirectangular relative position embedding; DAS, Distance-based attention scoring; EaAR, Equirectangular-aware attention rearrangement

Baseline transformer-Unet transformer using the Wilcoxon Signed-Rank test, where *p*-values less than 0.05 were considered statistically significant).

The baseline transformer-Unet transformer (B) achieved a mean accuracy of 84.56%. The sequential integration of our components consistently led to performance improvements. Adding the ERPE to the baseline (B + ERPE) significantly increased the mean accuracy to 90.42% ($p<0.001$ vs. B). This demonstrates ERPE's crucial role in mitigating geometric distortions by incorporating spherical positional information.

Further enhancing the model with DAS in the B + ERPE + DAS configuration resulted in an even higher mean accuracy of 93.74% ($p<0.001$ vs. B). This improvement highlights the effectiveness of DAS in prioritizing geometrically significant regions, thereby focusing attention more effectively.

Finally, the complete proposed model, integrating EaAR (B + ERPE + DAS + EaAR), achieved the highest mean accuracy of 94.79% ($p<0.001$ vs. B). This cumulative improvement from the baseline represents an absolute increase of approximately 10.23 percentage points, underscoring the synergistic benefits of all three components.

Notably, the full model (B + ERPE + DAS + EaAR) also exhibits the lowest standard deviation (0.31), indicating superior robustness and consistency across multiple runs compared to the baseline and intermediate configurations. These results unequivocally demonstrate that each proposed component contributes significantly to the enhanced accuracy and reliability of our geometry-biased transformer for dental caries detection.

To address the vital aspect of model explainability and interpretability in healthcare diagnostics, and to qualitatively demonstrate the contribution of our proposed modules (ERPE, DAS, EaAR) to improved spatial localization of caries lesions, we utilized Grad-CAM. Figure 8 illustrates the Grad-CAM outputs for a representative panoramic dental image from the CariesXrays dataset. Figure 7A presents the original panoramic image with ground-truth bounding boxes for dental caries, annotated by an expert, serving as our reference. Figure 7B illustrates the Grad-CAM output produced by the baseline transformer model, which mistakenly classifies the attrition facet of an additional tooth as a carious lesion, in addition to the true lesions, thereby revealing limitations in precise localization. Upon integrating the ERPE module, as shown in Fig. 7C, the issue of the falsely detected carious tooth persists, suggesting that while relative positional information is beneficial, it is insufficient to fully resolve localization challenges. Furthermore, Fig. 7D illustrates the Grad-CAM output after incorporating DAS alongside ERPE; the erroneous detection of an extra carious tooth is still present, highlighting the complexity of accurately localizing small lesions within the wide panoramic field of view. Crucially, Fig. 7E presents the Grad-CAM output after the complete integration of all proposed modules: ERPE, DAS, and EaAR. In this final configuration, the model's attention is precisely focused only on the actual caries lesions, and the false positive observed in previous stages is eliminated. This qualitative result strongly demonstrates that the combined effect

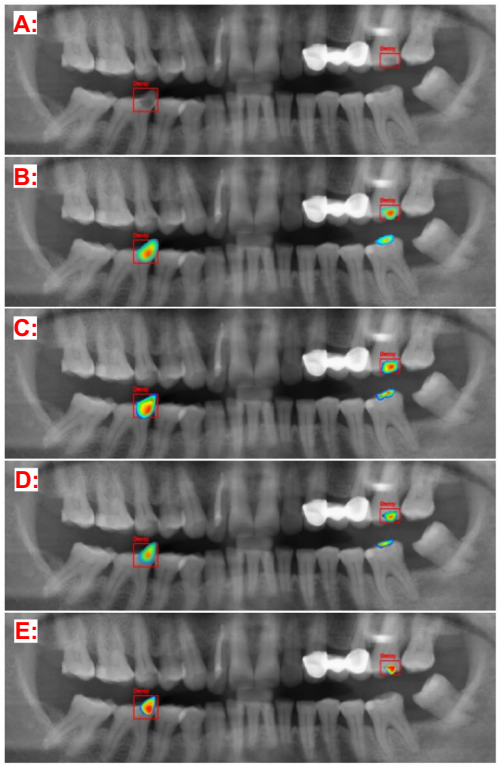

Fig. 7 Grad-CAM visual attention maps for a panoramic dental image from the CariesXrays dataset, demonstrating improved spatial localization of caries lesions with proposed modules. **A**: Original image with expert-annotated caries bounding boxes. **B**: Baseline transformer model's Grad-CAM. **C**: With ERPE **D**: With DAS added to ERPE **E**: With EaAR, ERPE, and DAS

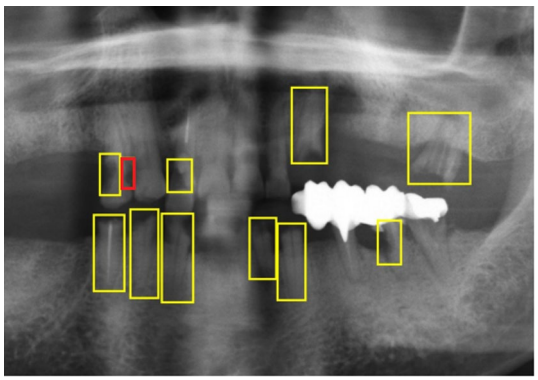

Fig. 8 Example of a missed subtle proximal caries (red box) on the CariesXrays dataset due to low radiographic contrast and overlapping of adjacent teeth

of our proposed modules significantly enhances the model's ability to achieve accurate spatial localization of caries and improves its interpretability by providing clear and correct attention maps, underscoring the importance of tailored architectural designs for medical image analysis.

Although the geometry-biased transformer demonstrates strong cross-dataset generalizability, we acknowledge that real clinical workflow validation remains outside the scope of this study. Our current experiments are conducted on multiple publicly available datasets under controlled conditions, which allows systematic comparison but does not fully reflect the variability of real clinical environments.

Furthermore, while our quantitative results are comprehensive, we did not include a dedicated qualitative investigation of failure cases that commonly arise in clinical practice. Examples of scenarios that may challenge the model in real-world settings include:

- **Subtle or Early-Stage Caries:** Very small incipient lesions with limited radiolucency may require finer-grained spatial feature modeling than what is available in current datasets.
- **Superimposed or Overlapping Structures:** Anatomical overlaps, restorations, and the inherent 3D-to-2D projection of panoramic imaging can obscure lesions, potentially reducing detection accuracy.
- **Low-Quality or Poorly Exposed Images:** Noise, motion artifacts, and contrast inconsistencies–frequently encountered in routine practice–may affect performance despite the robustness provided by the geometry-biased design.
- **Irregular or Atypical Lesion Morphology:** Complex caries shapes that deviate from common patterns may be misclassified due to limited representation in benchmark datasets.

For a concrete visual example of these limitations, Fig. 8 shows a real test image from the CariesXrays dataset in which the model failed to detect a subtle proximal caries (indicated by the red box). The lesion is characterized by extremely low radiographic contrast and is located in an area of significant tooth overlap, representing a combination of the two most common failure modes discussed above: subtle/early-stage caries and superimposed anatomical structures. This case illustrates that, despite the robustness provided by the geometry-biased design, certain clinically challenging lesions still require clinician verification.

Validating the model across these conditions in a full end-to-end clinical workflow, including acquisition variability, clinician–AI interaction, and integration into diagnostic routines, represents an essential direction for future work.

Beyond algorithmic performance, seamless integration into clinical Picture Archiving and Communication Systems (PACS) remains a key challenge. Although our model achieves low computational cost and sub-second inference times suitable for real-time use, practical deployment requires full "Digital Imaging and Communications in Medicine" compatibility, secure handling of patient metadata, compliance with "Health Insurance Portability and Accountability Act/General Data Protection" regulations, and interoperability with existing radiological workstations. Future work will therefore focus on developing a PACS-integrated prototype with interactive explainability features (e.g., direct overlay of Grad-CAM heatmaps) and conducting pilot studies in dental hospitals to evaluate end-to-end usability and clinical impact.

6 Conclusions

This study introduced a geometry-biased transformer to enhance dental caries detection in panoramic X-ray images by effectively addressing inherent geometric distortions through equirectangular relative position embedding, distance-based attention scoring, and equirectangular-aware attention rearrangement. Our model's ability to accurately model spherical geometry significantly improves spatial feature representation for both local and global dental structures. For evaluation, our primary assessment on the large-scale CariesXrays dataset demonstrated exceptional performance, achieving 94.90% accuracy and 0.9603 AUC, significantly outperforming existing methods. Crucially, to validate broader applicability and generalizability, we conducted comprehensive comparative analyses on several other public dental imaging datasets, including MICCAI STS-2D Challenge Dataset, Dental Radiography Dataset, and Panoramic Dental Dataset. The consistent and competitive performance across these diverse datasets highlights our method's versatility and potential for robust application beyond a single data source. In conclusion, the Geometry-Biased transformer not only enhances caries detection precision but also lays a strong foundation for geometry-aware transformers in medical image analysis. Future work will focus on further validating generalizability across an even wider range of clinical datasets from diverse settings and, critically, on the rigorous testing and integration of our prototype into real-world clinical workflows. This includes pilot studies and close collaboration with dental professionals to ensure clinical viability, addressing practical challenges like system integration, user interface design, real-time performance, and regulatory compliance, ultimately contributing to improved patient care outcomes.

Author contributions Nima Esmi and Koosha Refahi developed the methodology and conducted the experiments. Asadollah Shahbahrami supervised the technical aspects. Negar Khosravifard provided domain expertise in dental radiology and interpretation. Georgi Gaydadjiev contributed to the critical revision of the manuscript. All authors read and approved the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability Data will be made available on request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Jevdjevic M, Listl S (2025) Global, regional, and country-level economic impacts of oral conditions in 2019. J Dent Res 104(1):17–21. https://doi.org/10.1177/00220345241281698
2. Jin P, Wang L, Chen D, Chen Y (2024) Unveiling the complexity of early childhood caries: Candida albicans and Streptococcus mutans cooperative strategies. J Oral Microbiol 16(1):2339161. https://doi.org/10.1080/20002297.2024.2339161
3. Rokhshad R, Nasiri F, Saberi N, Shoorgashti R, Ehsani SS, Nasiri Z et al (2025) Deep learning for age estimation from panoramic radiographs: a systematic review and meta-analysis. J Dent 154:105560. https://doi.org/10.1016/j.jdent.2025.105560
4. Ver Berne J, Baseri Saadi S, Oliveira-Santos N, Marinho-Vieira LE, Jacobs R (2025) Automated classification of panoramic radiographs with inflammatory periapical lesions using a CNN-LSTM architecture. J Dent 156:105688. https://doi.org/10.1016/j.jdent.2025.105688
5. Chen Q, Zhao Y, Liu Y, Sun Y, Yang C, Li P et al (2021) MSLPNet: multi-scale location perception network for dental panoramic X-ray image segmentation. Neural Comput Appl 33(16):10277–10291. https://doi.org/10.1007/s00521-021-05790-5
6. Lin S, Hao X, Liu Y, Yan D, Liu J, Zhong M (2023) Lightweight deep learning methods for panoramic dental X-ray image segmentation. Neural Comput Appl 35(11):8295–8306. https://doi.org/10.1007/s00521-022-08102-7
7. Zhu H, Cao Z, Lian L, Ye G, Gao H, Wu J (2023) CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image. Neural Comput Appl 35(22):16051–16059. https://doi.org/10.1007/s00521-021-06684-2
8. Haghanifar A, Majdabadi MM, Haghanifar S, Choi Y, Ko SB (2023) PaXNet: tooth segmentation and dental caries detection in panoramic X-ray using ensemble transfer learning and capsule classifier. Multimed Tools Appl 82(18):27659–27679. https://doi.org/10.1007/s11042-023-14435-9
9. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE, Mahmoudinia E et al (2022) Deep learning for caries detection: a systematic review. J Dent 122:104115. https://doi.org/10.1016/j.jdent.2022.104115
10. Son JY, Park Y, Park JY, Kim MJ, Han DH (2024) Overdiagnosis of dental caries in South Korea: a pseudo-patient study. BMC Oral Health 24(1):1–10. https://doi.org/10.1186/s12903-024-05061-4
11. Alharbi SS, Alhasson HF (2024) Exploring the applications of artificial intelligence in dental image detection: a systematic review. Diagnostics 14(21):2442. https://doi.org/10.3390/diagnostics14212442

12. Sankaran KS (2022) An improved multipath residual CNN-based classification approach for periapical disease prediction and diagnosis in dental radiography. Neural Comput Appl 34(22):20067–20082. https://doi.org/10.1007/s00521-022-07556-z
13. Can Z, Isik S, Anagun Y (2024) CVApool: using null-space of CNN weights for the tooth disease classification. Neural Comput Appl 36(26):16567–16579. https://doi.org/10.1007/s00521-024-09995-2
14. Park EY, Cho H, Kang S, Jeong S, Kim EK (2022) Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. BMC Oral Health 22(1):573. https://doi.org/10.1186/s12903-022-02589-1
15. Roy M, Baruah U, Varma V (2024) TransDL: a transfer learning-based concatenated model for Covid-19 identification and analysis of posteroanterior chest X-ray images. Multimed Tools Appl 83(11):33421–33443. https://doi.org/10.1007/s11042-023-16825-5
16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. Preprint at arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
17. Yang X, Li X, Li X, Wu P, Shen L, Deng Y (2024) ImplantFormer: vision transformer-based implant position regression using dental CBCT data. Neural Comput Appl 36(12):6643–6658. https://doi.org/10.1007/s00521-023-09411-1
18. Esmi N, Shahbahrami A, Gaydadjiev G, de Jonge P (2025) Suicide ideation detection based on documents dimensionality expansion. Comput Biol Med 192:110266. https://doi.org/10.1016/j.compbiomed.2025.110266
19. Aburass S, Dorgham O, Al Shaqsi J, Abu Rumman M, Al-Kadi O (2025) Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. J Imaging Inform Med. https://doi.org/10.1007/s10278-025-01481-y
20. Esmi N, Shahbahrami A, Gaydadjiev G, de Jonge P (2025) TEREE: transformer-based emotion recognition using EEG and Eye movement data. Intell Based Med 12:100305. https://doi.org/10.1016/j.ibmed.2025.100305
21. Schneider L, Krasowski A, Pitchika V, Bombeck L, Schwendicke F, Büttner M (2025) Assessment of CNNs, transformers, and hybrid architectures in dental image segmentation. J Dent 156:105668. https://doi.org/10.1016/j.jdent.2025.105668
22. Hao J, Wong LM, Shan Z, Ai QYH, Shi X, Tsoi JKH et al (2024) A semi-supervised transformer-based deep learning framework for automated tooth segmentation and identification on panoramic radiographs. Diagnostics 14(17):1948. https://doi.org/10.3390/diagnostics14171948
23. Yun I, Shin C, Lee H, Lee HJ, Rhee CE (2023) Egformer: equirectangular geometry-biased transformer for 360 depth estimation. In: Conference on computer vision and pattern recognition. pp 6101–6112
24. Chen B, Fu S, Liu Y, Pan J, Lu G, Zhang Z (2024) CariesXrays: enhancing caries detection in hospital-scale panoramic dental X-rays via feature pyramid contrastive learning. In: AAAI conference on artificial intelligence. pp 21940–21948
25. Li KC, Mao YC, Lin MF, Li YQ, Chen CA, Chen TY et al (2024) Detection of tooth position by YOLOv4 and various dental problems based on CNN With bitewing radiograph. IEEE Access 12:11822–11835. https://doi.org/10.1109/ACCESS.2023.3348788
26. Cotti E, Schirru E (2022) Present status and future directions: imaging techniques for the detection of periapical lesions. Int Endod J 55:1085–1099. https://doi.org/10.1111/iej.13828
27. DeWood G (2017) Occlusal Bite Disease. Spear Education
28. Gupta A (2023) On imaging modalities for cephalometric analysis: a review. Multimed Tools Appl 82(24):36837–36858. https://doi.org/10.1007/s11042-023-14971-4
29. Xue T, Chen L, Sun Q (2024) Deep learning method to automatically diagnose periodontal bone loss and periodontitis stage in dental panoramic radiograph. J Dent 150:105373. https://doi.org/10.1016/j.jdent.2024.105373
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. In: Advances in neural information processing systems. pp 1–15. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
31. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Conference on computer vision and pattern recognition. pp 10012–10022
32. Sener E, Onem E, Akar GC, Govsa F, Ozer MA, Pinar Y et al (2018) Anatomical landmarks of mandibular interforaminal region related to dental implant placement with 3D CBCT: comparison between edentulous and dental mandibles. Surg Radiol Anat 40:615–623. https://doi.org/10.1007/s00276-017-1934-8
33. Ryu J, Lee DM, Jung YH, Kwon O, Park S, Hwang J et al (2023) Automated detection of periodontal bone loss using deep learning and panoramic radiographs: a convolutional neural network approach. Appl Sci 13(9):5261. https://doi.org/10.3390/app13095261
34. Szabó V, Orhan K, Dobó-Nagy C, Veres DS, Manulis D, Ezhov M et al (2025) Deep learning-based periapical lesion detection on panoramic radiographs. Diagnostics 15(4):510. https://doi.org/10.3390/diagnostics15040510
35. Singh SB, Laishram A, Thongam K, Singh KM (2025) Automatic detection and classification of dental anomalies and tooth types using transformer-based Yolo with GA optimization. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3556523
36. Anbarasi LJ, Neeraja R, Sharen H, Jawahar M, Ravi V (2024) An enhanced caries detection and prediction using DentSU_Net. In: IoT Sensors, ML, AI and XAI: empowering a smarter world. pp 439–454

37. Ying S, Huang F, Shen X, Liu W, He F (2024) Performance comparison of multifarious deep networks on caries detection with tooth X-ray images. J Dent 144:104970. https://doi.org/10.1016/j.jdent.2024.104970
38. Ayhan B, Ayan E, Atsü S (2025) Detection of dental caries under fixed dental prostheses by analyzing digital panoramic radiographs with artificial intelligence algorithms based on deep learning methods. BMC Oral Health 25(1):216. https://doi.org/10.1186/s12903-025-05577-3
39. van Nistelrooij N, El Ghoul K, Xi T, Saha A, Kempers S, Cenci M et al (2024) Combining public datasets for automated tooth assessment in panoramic radiographs. BMC Oral Health 24(1):387. https://doi.org/10.1186/s12903-024-04129-5
40. Sun D, Wang J, Zuo Z, Jia Y, Wang Y (2024) STS-TransUNet: semi-supervised tooth segmentation transformer U-Net for dental panoramic image. Math Biosci Eng 21(2):2366–2384. https://doi.org/10.3934/mbe.2024104
41. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision. pp 213–229
42. Küçük DB, Imak A, Özçelik STA, Çelebi A, Türkoğlu M, Sengur A (2025) Hybrid CNN-transformer model for accurate impacted tooth detection in panoramic radiographs. Diagnostics 15(3):244. https://doi.org/10.3390/diagnostics15030244
43. Alsakar YM, Elazab N, Nader N, Mohamed W, Ezzat M, Elmogy M (2024) Multi-label dental disorder diagnosis based on MobileNetV2 and swin transformer using bagging ensemble classifier. Sci Rep 14(1):25193. https://doi.org/10.1038/s41598-024-73297-9
44. Kim MJ, Chae SG, Bae SJ, Hwang KG (2024) Unsupervised few-shot learning architecture for diagnosis of periodontal disease in dental panoramic radiographs. Sci Rep 14(1):23237. https://doi.org/10.1038/s41598-024-73665-5
45. Cho S, Jung R, Kwon J (2024) Sampling based spherical transformer for 360 degree image classification. Expert Syst Appl 238:121853. https://doi.org/10.1016/j.eswa.2023.121853
46. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q et al (2022) Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp 205–218
47. Ling Z, Xing Z, Zhou X, Cao M, Zhou G (2023) Panoswin: a pano-style Swin transformer for panorama understanding. In: Conference on computer vision and pattern recognition. pp 17755–17764
48. MICCAI STS Challenge Organizers (2024) STS MICCAI 2024 challenge: grand challenge on 2D and 3D semi-supervised tooth segmentation. MICCAI STS Challenge
49. Imtkaggleteam (2023) Dental radiography dataset. Kaggle
50. Thunderpede (2023) Panoramic Dental Dataset. Kaggle
51. Isbrycee (2024) TSI15k: Semi-Supervised Tooth Segmentation Dataset. GitHub
52. Hamamci IE, Er S, Simsar E, Yuksel AE, Gultekin S, Ozdemir SD et al (2023) DENTEX: an abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic X-rays. pp 1–6. https://doi.org/10.48550/arXiv.2305.19112

Authors and Affiliations

Nima Esmi[1,2] · Koosha Refahi[3] · Asadollah Shahbahrami[2,3] · Negar Khosravifard[4] · Georgi Gaydadjiev[5]

✉ Nima Esmi
n.esmi.rudbardeh@rug.nl

Koosha Refahi
koosharefahi@gmail.com

Asadollah Shahbahrami
shahbahrami@guilan.ac.ir

Negar Khosravifard
negarkhosravifard@gums.ac.ir

Georgi Gaydadjiev
G.N.Gaydadjiev@tudelft.nl

[1] Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Groningen University, Groningen, The Netherlands

[2] Intelligent Systems Research Center, Khazar University, Baku, Azerbaijan

[3] Department of Computer Engineering, University of Guilan, Guilan, Iran

[4] Department of Oral and Maxillofacial Radiology, Dental Sciences Research Center, School of Dentistry, Guilan University of Medical Sciences, Rasht, Guilan, Iran

[5] Computer Engineering Laboratory, Delft University of Technology, Delft, The Netherlands