technologies

MDPI

Article

A Dual-Stream Dental Panoramic X-Ray Image Segmentation Method Based on Transformer Heterogeneous Feature Complementation

Tian Ma, Jiahui Li *, Zhenrui Dang, Yawen Li and Yuancheng Li

College of Artificial Intelligence and Computer Science, Xi'an University of Science and Technology, Xi'an 710054, China

* Correspondence: 23208088015@stu.xust.edu.cn

Abstract

To address the widespread challenges of significant multi-category dental morphological variations and interference from overlapping anatomical structures in panoramic dental X-ray images, this paper proposes a dual-stream dental segmentation model based on Transformer heterogeneous feature complementarity. Firstly, we construct a parallel architecture comprising a Transformer semantic parsing branch and a Convolutional Neural Network (CNN) detail capturing pathway, achieving collaborative optimization of global context modeling and local feature extraction. Furthermore, a Pooling-Cooperative Convolutional Module was designed, which enhances the model's capability in detail extraction and boundary localization through weighted centroid features of dental structures and a latent edge extraction module. Finally, a Semantic Transformation Module and Interactive Fusion Module are constructed. The Semantic Transformation Module converts geometric detail features extracted from the CNN branch into high-order semantic representations compatible with Transformer sequential processing paradigms, while the Interactive Fusion Module applies attention mechanisms to progressively fuse dual-stream features, thereby enhancing the model's capability in holistic dental feature extraction. Experimental results demonstrate that the proposed method achieves an IoU of 91.49% and a Dice coefficient of 94.54%, outperforming current segmentation methods across multiple evaluation metrics.

Keywords: medical image segmentation; tooth segmentation; CNN; transformer

check for updates

Academic Editors: Thomas Lindner and Manoj Gupta

Received: 23 May 2025
Revised: 17 June 2025
Accepted: 30 June 2025
Published: 8 July 2025

Citation: Ma, T.; Li, J.; Dang, Z.; Li, Y.; Li, Y. A Dual-Stream Dental Panoramic X-Ray Image Segmentation Method Based on Transformer Heterogeneous Feature Complementation. *Technologies* **2025**, *13*, 293. https://doi.org/10.3390/technologies13070293

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Dental panoramic X-ray imaging through arcuate tomographic scanning, achieves precise visualization of oral and maxillofacial composite anatomical structures. By including the complete details of teeth, jaws, and surrounding soft tissues, it offers vital diagnostic assistance for diseases surrounding the jaws. Panoramic images enable precise evaluation of dental inclination angles, periodontal soft tissues, roots, and alveolar bone status. Thus, they are widely used in dental morphology analysis, dental crowding analysis, molar analysis, alveolar bone morphology research, and implant surgical planning. To boost the analysis efficiency of panoramic X-ray medical images, researchers are exploring automated segmentation approaches. Lin et al. [1] proposed an automated segmentation method based on local singularity analysis for dental segmentation in periapical radiographs. Park et al. [2] systematically reviewed the evolution of dental image processing technologies, pointing out that deep learning is progressively replacing traditional image

processing methods to become the dominant technological paradigm. Research on these methods has improved the segmentation efficiency of dental panoramic X-ray images in clinical work. The segmented images offer key information for evaluating missing teeth, assessing tooth development, and analyzing the position of embedded teeth and their relationship with adjacent teeth.

Traditional dental segmentation methods mainly rely on image processing and machine learning techniques, which require manual feature extraction and classifier design. However, in complex scenarios, traditional methods struggle to accurately identify adjacent teeth when dealing with complex tooth shapes and overlapping anatomical structures. This is due to the limitations of manual feature extraction and the difficulty of classifiers in distinguishing various complex situations. As deep learning technology has advanced, particularly with the emergence of CNN, it has been widely applied in the field of image segmentation. Arora et al. [3] employed a multimodal CNN architecture, using original grayscale X-ray images and their corresponding edge maps as inputs. They leveraged the grayscale intensity information from the images and the structural features from the edge maps for dental segmentation. Hou et al. [4] proposed Teeth U-Net, a segmentation model for contextual semantics and contrast enhancement in dental panoramic X-ray images, which improves segmentation accuracy by enhancing contrast to accentuate the disparity between teeth and background structures. Chandrashekar et al. [5] developed a collaborative deep learning model based on panoramic X-rays for dental segmentation and identification.Compared with traditional methods, the above CNN-based dental segmentation methods can automatically extract image features without manual intervention, so they are more robust and accurate. However, due to the similar morphological features of adjacent teeth, achieving accurate individual identification between them still poses challenges. CNN-based tooth segmentation algorithms, constrained by local receptive fields, exhibit inherent limitations in modeling long-range dependencies, thereby failing to effectively capture global contextual information within images and consequently encountering challenges in accurately classifying adjacent teeth. Therefore, the extraction of spatial prior information plays a crucial role in modeling dental morphology, tooth arrangement, and the relative positions of teeth [6,7]. In recent years, Transformer-based methods [8] have demonstrated promising segmentation performance in dental panoramic X-ray images by leveraging self-attention mechanisms to capture long-range contextual information. However, the self-attention mechanism in Transformer focuses on relationships between pixel-level tokens, ignoring intrinsic attributes like dental morphological structure and spatial location. This causes a topological integrity deficiency in its global context modeling.

Based on the above analysis, this paper proposes a dual-stream dental segmentation network with heterogeneous feature complementarity that integrates the morphological features and spatial prior information of teeth into the network framework to improve the model's accuracy and robustness. The main contributions are as follows:

1. To address the challenge of effectively coordinating global contextual information and local detailed features in tooth segmentation tasks, we constructed a dual-stream network architecture comprising a parallel Transformer-based semantic parsing branch and a CNN-based detail capturing path. The Transformer branch captures global contextual information and long-range dependencies among pixels, while the CNN branch provides rich local details for precise dental segmentation. This global-local complementary mechanism effectively achieves collaborative optimization of global context modeling and local feature extraction, fully leveraging the strengths of both approaches to enhance the overall performance of tooth segmentation.

2. To address the challenges posed by the complex spatial distribution and diverse morphological details in dental anatomical structures, this paper proposes a Pooling-Cooperative Convolutional Module (PCM). The module employs three pooling operations with weighted feature weighting to represent dental centroids, thereby enhancing the spatial feature representation capability of CNN branches and improving their detail extraction performance. A latent edge extraction module is designed to enhance the synergistic representation of periodontal boundary features and internal textural features, thereby improving the accuracy of boundary segmentation.
3. To address the modality heterogeneity between CNN and Transformer, a Semantic Transformation Module (STM) is designed, which performs semantic space mapping on CNN-extracted localized detail features to reconfigure them into high-dimensional semantic information compatible with Transformer architectures. An Interactive Fusion Module (IFM) is designed to achieve multi-scale feature fusion from Transformer and CNN, ensuring that the fused features simultaneously retain global contextual dependencies and local fine-grained information, thereby enhancing the model's recognition accuracy for low-resolution semantic features and its representational capacity for complex spatial information.

2. Related Work

Dental segmentation has emerged as a fundamental task in the medical image processing domain. The current segmentation methodologies are primarily categorized into two classes: traditional methodologies based on raw image features and deep learning-based approaches driven by deep image features. Traditional segmentation methods mainly rely on the intrinsic characteristics of dental panoramic X-rays (e.g., shape, grayscale intensity) for tooth segmentation, including watershed-based segmentation algorithms [9], threshold-based segmentation algorithms [10], clustering-based segmentation algorithms [11], boundary-based segmentation algorithms [12], and region-growing-based segmentation algorithms [13]. However, the aforementioned traditional segmentation methods commonly exhibit sensitivity to image grayscale intensity, noise, and initial conditions when processing complex images, resulting in constrained segmentation accuracy and the poor stability of segmentation results. These methods predominantly rely on singular image features (e.g., grayscale intensity) and often struggle to accurately differentiate target structures from background tissues when encountering images with low grayscale contrast, excessive noise, or complex backgrounds, frequently leading to over-segmentation or under-segmentation artifacts. In recent years, deep learning techniques have progressively superseded traditional methodologies due to their superior performance, demonstrating extensive applications in the medical image segmentation and disease detection domains.

The first category encompasses CNN-based approaches. PaXNet, proposed by Haghanifar et al. [14], employs a multi-task learning strategy that enables feature information sharing between dual tasks and is specifically designed to address technical challenges such as substantial morphological variations in teeth and ill-defined caries boundaries within panoramic radiographs. Arora et al. [3] developed a multi-modal CNN architecture achieving automated segmentation on dental panoramic X-rays. Tekin et al. [15] implemented a Mask R-CNN-based approach for simultaneous tooth segmentation and numbering in X-ray images, attaining high-accuracy segmentation outcomes. The U-Net architecture proposed by Ronneberger et al. [16] has been extensively adopted in the medical image segmentation domain due to its simple yet effective architecture and superior performance, establishing itself as a prevalent segmentation framework in medical image processing. With the widespread application of U-Net, different researchers have explored various structural improvements based on this model. Koch et al. [17] implemented

the semantic segmentation of dental panoramic images based on the U-Net architecture, where their model relied on a smaller and simpler network architecture and demonstrated excellent performance in sample segmentation. Zhao et al. [18] proposed a Two-Stage Attention Segmentation Network (TSASNet) to address the challenges of blurred tooth boundaries and difficulties in root segmentation. This network integrates attention modules and CNN architectures, which are respectively designed for localizing dental regions and precisely segmenting actual tooth regions from attention maps. Hou et al. [4] proposed Teeth U-Net, which effectively extracts the contextual feature information of teeth through a dense skip connection mechanism and multi-scale aggregated attention blocks. Additionally, Ma et al. [19] proposed an improved ICNet architecture integrated with attention mechanisms for dental lesion segmentation. By introducing a lightweight Convolutional Block Attention Module (CBAM), this study enhances the capability to learn discriminative features of dental pathologies, particularly in complex scenarios where lesions exhibit high morphological similarity to healthy tissues. This method not only achieves enhanced segmentation accuracy and real-time computational performance but also provides novel technical insights for advancing automated diagnosis of dental diseases.

Although convolutional neural networks demonstrate superior performance in small-scale image analysis tasks owing to their inductive biases, the intrinsic local receptive fields of convolutional operations constrain the model's capacity for a global semantic relationship. This limitation may result in critical contextual information loss when handling complex semantic scenarios. To address this limitation, TransUNet [20] established a new paradigm in medical image segmentation by integrating the global encoding capabilities of Transformers with the upsampling functionalities of U-Net architectures. Li et al. [21] proposed GT U-Net, which introduces enhanced-performance Group Transformers to replace conventional encoder–decoder architectures, while reducing computational complexity through integrated grouping structures and bottleneck mechanisms. Unter [22] further transforms the volumetric medical image segmentation task into a sequence prediction problem, signifying a significant advancement in the application of Transformer architectures to 3D medical imaging. Swin-Unet [23] integrates the topological architecture of U-Net with the attention mechanism from Swin Transformer, while employing patch expanding layers for upsampling operations. This hybrid architecture demonstrates superior performance on multi-organ CT segmentation and the ACDC dataset. Furthermore, the Mask-Transformer-based architecture [24] has demonstrated significant performance advantages in tooth segmentation tasks. This algorithm employs a dual-path design and incorporates a panoramic quality loss function, which simplifies the training process. Although these methods fully leverage the global dependency modeling capabilities of Transformer encoders, they often overemphasize global feature extraction while neglecting effective integration of local features. To address the research challenge of synergistic optimization between global semantic modeling and local detail preservation in existing methods, Ma et al. [25] proposed a multi-feature coordinate position learning framework. Through the collaborative design of Residual Omnidirectional Dynamic Convolution (ROCM) and Two-Stream Coordinate Attention (TSCA) this approach achieves effective fusion of local features with global contextual information. However, this method still has room for improvement in cross-scale feature interaction and the quantitative modeling of spatial relationships. Based on the above analysis, this paper proposes a two-stream network architecture integrating CNN and Transformer to enhance the analytical capacity for dental feature focusing and their spatial topological correlations.

3. Research Design

3.1. Overall Architecture

This paper proposes a heterogeneous feature-complementary dual-stream network architecture that integrates the complementary advantages of Transformer and CNN, constructing a dual-path feature extraction network with global-local collaborative perception capabilities. As illustrated in Figure 1, the model comprises two parallel branches: a Transformer-based spatial perception branch and a CNN-based semantic feature extraction branch. The spatial perception branch employs the Transformer architecture, leveraging its self-attention mechanism and global contextual modeling capability to establish long-range dependency relationships across pixels. The CNN semantic branch extracts localized detail features from dental medical images through its convolutional neural network architecture. Finally, the features extracted from the two branches are fused through weighted integration, which not only enhances the spatial localization accuracy of feature maps in critical anatomical details (e.g., alveolar bone boundaries and occlusal surfaces of tooth crowns) but also preserves the semantic completeness required for dental tissue classification and pathological region differentiation. This complementary synergy mechanism fully capitalizes on the architectural strengths of dual-branch networks, effectively addressing mis-segmentation challenges in dental panoramic X-ray image segmentation caused by global contextual insufficiency or localized feature ambiguity.

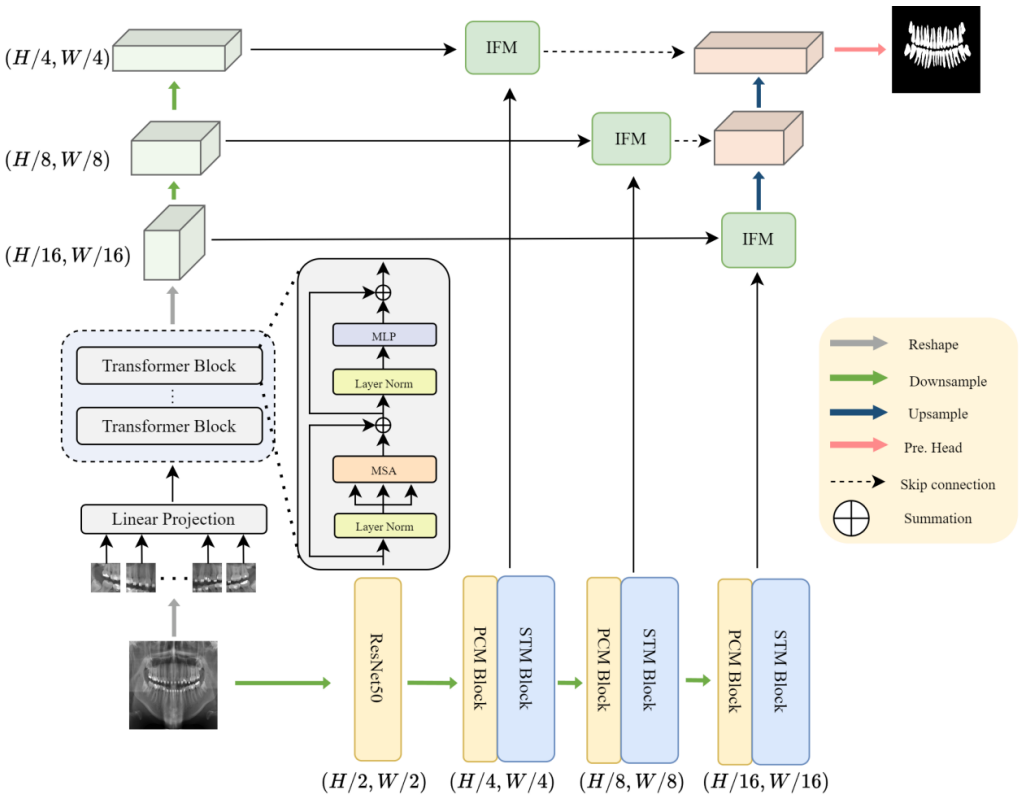


Figure 1. Overall network architecture diagram.

Specifically, in the encoder section, the network incorporates a parallel dual-encoder architecture composed of a Transformer branch and a CNN branch. To optimize model parameters and enhance segmentation accuracy for dental X-ray images, the Transformer branch employs a ViT-B16 model pretrained on ImageNet as its backbone. This hierarchically stacked six-layer Transformer establishes pixel-level global correlations through multi-head self-attention mechanisms, thereby augmenting the feature representational capacity of the model in handling complex imaging data. For local feature extraction,

the network employs ResNet50 as the backbone of the CNN branch, where its residual connection architecture with cross-layer mapping effectively mitigates gradient vanishing issues. To further enhance the network's representational capacity for localized detail features, a Pooling-Cooperative Convolutional Module is embedded within the CNN branch. This module preserves the integrity of downsampling pathways while significantly improving the capture capability for fine-grained anatomical characteristics, such as enamel micro-textures and root morphological variations, thereby providing rich and precise local information to subsequent image segmentation tasks. Through a Semantic Transformation Module, the local extracted features from the CNN branch are transformed into transformer-compatible semantic representations, enabling effective fusion of cross-branch features. Furthermore, in the skip connection phase, an Interactive Fusion Module is introduced that employs a channel-spatial dual attention weighting mechanism to the enhance information interaction between the global contextual features from the Transformer branch and the local high-resolution features derived from the CNN branch. Finally, features with identical resolutions are subsequently fused and the integrated features are then propagated to the decoder. Finally, the features sharing identical resolutions undergo fusion, and the resultant fused features are propagated to the decoder for subsequent processing and image restoration. This mechanism ensures the accuracy of final segmentation outcomes while preserving detailed representations, thereby effectively addressing the multi-scale feature extraction requirements to dental image segmentation. The subsequent sections will systematically elaborate on the operational principles and implementation mechanisms of each constituent module within the network architecture.

3.2. Pooling-Cooperative Convolutional Module

Building upon the aforementioned architecture, a Pooling-Cooperative Convolutional Module is designed to effectively address the limitations of existing methods in capturing both the positional features of targets and edge-related characteristics. As illustrated in Figure 2, the PCM module enhances the model's capability in perceiving target locations and capturing edge details by combining three pooling strategies (max pooling, average pooling, and median pooling) to achieve diversified feature representations, while collaboratively integrating positional features across both channel and spatial dimensions. It specifically proceeds as follows:

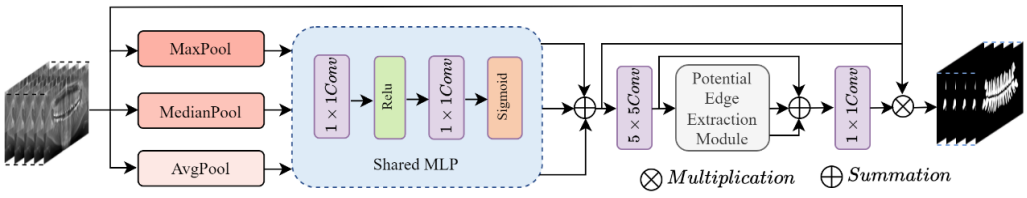


Figure 2. Pooling-Cooperative Convolutional Module.

The input feature map is first processed through three global pooling operations: global average pooling (AvgPool), global max pooling (MaxPool), and global median pooling (MedianPool). Since all three pooling operations are global, the size of each pooling operation matches the spatial dimensions of the input feature map. Global average pooling computes spatial mean values across all positions per feature map channel, generating channel-wise feature vectors to extract holistic average information. This approach preserves the overall grayscale distribution trend of teeth, effectively captures the global representation of jawbone and tooth morphology, and mitigates interference from local noise. Global max pooling focuses on salient features by preserving the most prominent features within feature maps. This method is particularly suitable for capturing structures

with pronounced radiodensity variations, such as tooth boundaries and dental caries, thereby enhancing the model's sensitivity to salient features. Global median pooling derives the median values per channel, exhibiting robustness against outliers while extracting stable feature representations. It is therefore robust to common dental imaging artifacts such as metal artifacts and scanning noise and can extract stable background and tooth transition features. Other pooling techniques—such as minimum pooling, which focuses on low-density regions—tend to discard the high-density features of the dental structure, consequently leading to incomplete segmentation areas. Stochastic pooling introduces training randomness through random sampling, yet exhibits poor stability during the testing phase, failing to meet the consistency requirements for medical image segmentation. The three pooling operations yield three distinct pooling results, each with a dimension of $\mathbb{R}^{C\times1\times1}$ (where C denotes the number of channels).

Subsequently, the features from each pooling branch are fed into a parameter-shared multilayer perceptron (MLP), as formulated in Equation (1). The MLP architecture comprises two 1 × 1 convolutional layers, ReLU and Sigmoid activation functions. Within the MLP, the first convolutional layer performs channel-wise dimension compression, compressing feature channels from C to C/r (where r denotes the reduction ratio), effectively mitigating feature redundancy. The processed features are subsequently passed through a ReLU activation function to introduce nonlinear representational capacity, followed by restoration to the original channel dimension C via the second 1 × 1 convolutional layer, thereby preserving the integrity of the feature space. Ultimately, normalization is achieved through a Sigmoid activation function that constrains output values to the [0, 1] range, yielding three attention response maps. The attention maps generated by the three pooling strategies are aggregated via element-wise summation to produce a comprehensive channel attention map. Subsequently, this channel attention map is element-wise multiplied with the original input feature map to achieve dynamic feature calibration along the channel dimension, as detailed in Equation (2).

$$F_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F))) + \sigma(\mathrm{MLP}(\mathrm{MaxPool}(F))) + \sigma(\mathrm{MLP}(\mathrm{MedianPool}(F))), \quad (1)$$

$$F' = F_c \odot F, \quad (2)$$

where σ denotes the Sigmoid function and $\odot$ represents element-wise multiplication.

After completing the feature calibration along the channel dimension, the obtained comprehensive channel attention map is fed into a 5 × 5 depthwise convolutional layer to extract fundamental features. The convolutional layer maintains the same output size as the input, ensuring that the spatial information of the feature maps is preserved without loss. The output feature map is then processed through the Latent Edge Extraction Module to further extract edge information.

The implementation details of the Latent Edge Extraction Module are illustrated in Figure 3, with its core workflow comprising geometric feature decoupling and topology guidance. First, a 1 × 1 convolutional layer performs nonlinear mapping on deep features to generate the initial segmentation mask f_m. Subsequently, dual-path operations of morphological erosion and dilation are employed, and via element-wise subtraction operation $\ominus$, to construct a binary topological structure distinguishing boundary-internal regions. The erosion operation employs a 3 × 3 square structuring element, while the dilation operation utilizes a 3 × 3 cross-shaped structuring element. Specifically, during the topological decoupling stage, morphological erosion and dilation operations are applied to the initial segmentation mask. These operations enable dynamic evolution of edges through pixel-wise morphological iteration (with a single-pixel step size), thereby delineating the boundary and internal regions. This dual-region topological representation quantifies

the geometric confidence level of target regions, guiding subsequent anisotropic feature propagation in the feature refinement network. The separation process is formulated as

$$P_{CR} = E(f_m) \times T, \tag{3}$$

$$P_{BR} = (D(f_m) \times T - P_{CR}). \tag{4}$$

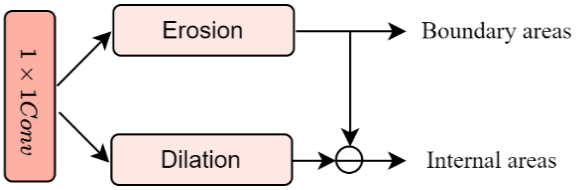


Figure 3. Latent Edge Extraction Module.

Herein, 1 denotes the internal region and 2 represents the potential edge region. The symbol × denotes matrix multiplication operation, T indicates the iteration count of the erosion operator, with the total operation counts for both erosion and dilation operators being identical. In the final feature fusion stage, multi-scale features output from all deep convolutional branches undergo cross-receptive field aggregation through element-wise summation operations, resulting in a multi-scale geometric feature map. A subsequent 1 × 1 convolutional layer then performs dimensionality reduction on the fused features, compacting redundant channel-wise information. The computed channel attention map is then element-wise multiplied by the channel-weighted feature maps to achieve differentiated enhancement of the feature channels while simultaneously suppressing irrelevant background noise, ultimately yielding a discriminatively enhanced feature map.

3.3. Semantic Transformation Module

In representation learning, CNN architectures and Transformer architectures exhibit inherent heterogeneity: the former constructs hierarchical local features through convolutional kernels' local receptive fields and translation invariance, while the latter employs self-attention mechanisms to achieve global long-range dependency modeling. To enable the CNN branch to effectively integrate high-quality semantic information from the Transformer branch, this paper proposes the STM, inspired by Zhao et al. [26]. The STM facilitates deep interaction of multi-dimensional features and the effective capture of cross-scale dependencies through multi-stage hierarchical processing.

As illustrated in the left panel of Figure 4, the STM adopts a Transformer-like topological architecture, incorporating core components including multi-head attention, batch normalization, and feed-forward networks (FFNs). The operational procedure can be formally described as follows:

$$f = \text{Norm}(x + \text{CondAttention}(x)), \tag{5}$$

$$y = \text{Norm}(f + \text{FFN}(f)), \tag{6}$$

where Norm denotes Batch Normalization, CondAttentation denotes Condensed Multi-Scale Attention, and x, f, and y represent the input, hidden features, and output, respectively.

Due to the significant semantic discrepancy between Transformers and CNNs, merely employing a similarity matching mechanism between learnable vectors and pixel-wise features can only establish locally correlated weight distribution maps. This unidimensional attention modeling paradigm exhibits systematic deficiencies in reconciling local textural sensitivity with global semantic integrity, consequently failing to sufficiently capture rich contextual information. To address the limitations of existing Transformer architectures in

multi-level semantic fusion and global relational modeling, this paper introduces a channel-space decoupled three-stage framework within the Semantic Transformation Module. As shown in the upper-right panel of Figure 4, the model achieves deep multi-dimensional feature interaction and effective capture of cross-scale dependencies through progressive processing comprising feature aggregation, attention computation, and feature restoration. This design preserves the structural integrity of features while the enhancing semantic representation capability.

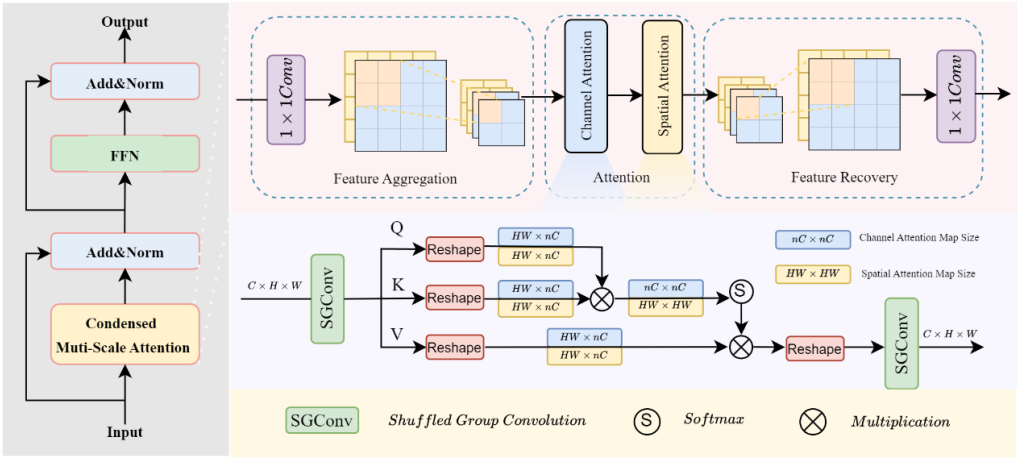


Figure 4. Semantic Transformation Module.

However, when integrating multi-head self-attention-based semantic representations from Transformer architectures, the pervasive high-dimensional redundant features in the channel and spatial domains significantly escalate computational complexity. Effectively compressing redundant features while recovering critical information thus constitutes a pivotal challenge. To address this, channel features and spatial features are first aggregated into condensed representations prior to attention computation, thereby achieving preliminary feature integration and refinement. Subsequently, to comprehensively capture the global dependencies of pixel features across both channel and spatial dimensions, this paper employs two serially-connected attention mechanisms that sequentially process pixel features through channel-wise attention followed by spatial-domain attention. The channel attention mechanism first models inter-channel feature correlations, while the spatial attention mechanism subsequently extracts cross-regional spatial long-range dependencies. Through this dual-dimensional paradigm, the model achieves the precise capture of global dependency relationships across both channel and spatial dimensions while effectively exploiting inter-feature correlations.

Given the input feature $F \in \mathbb{R}^{H\times W\times C}$, where $H \times W$ denotes the spatial resolution and C represents the number of channels, the channel attention mechanism aggregates features along the channel dimension, yielding a channel-condensed feature $\tilde{F} \in \mathbb{R}^{H\times W\times C'}$, where $C' = C/r_c.r_c$ are aggregation factors. The specific implementations are formally described in Equation (7):

$$\tilde{F} = \Phi^{CA}(F), \tag{7}$$

by introducing a pointwise convolution $\Phi^{CA}(\cdot)$ to adaptively aggregate informative features in the channel domain, the framework subsequently aggregates features along the spatial dimension to generate the spatially compressed feature $\hat{F} \in \mathbb{R}^{H'\times W'\times C''}$, where $H' = H/S$, $W' = W/S$, and $C'' = C'r_s$.

$$\hat{F} = \Psi^{SA}(\tilde{F}), \tag{8}$$

where $\Psi^{SA}(\cdot)$ aggregates spatial features of size $S \times S \times 1$ into compressed features of size $1 \times 1 \times r_s$, thereby generating feature $H/S \times W/S \times C'r_i$ with size $\hat{F}$. To reduce computational complexity, a grouped convolution is introduced with C' input channels and C'' output channels and employing kernel size $S \times S$ and a stride of S. Through this methodology, expanded channels enable comprehensive preservation and aggregation of spatial features, thereby capturing cross-domain feature information in both the channel and spatial domains while acquiring dependency relationships in low-dimensional spaces.

Following attention computation, to restore the feature distributions in the channel and spatial domains, the framework employs decoupled steps inverse to the aggregation process: first recovering spatial features, then restoring channel features. The mathematical formalization is articulated as follows:

$$\bar{F} = \Phi^{CR}(\Psi^{SR}(\Theta(\hat{F}))), \tag{9}$$

where $\Theta(\hat{F})$ denotes the post-attention features and $\Psi^{SR}(\cdot)$ and $\Phi^{CR}(\cdot)$ represent the spatial and channel feature restoration functions, respectively. $\Psi^{SR}(\cdot)$ employs a 1×1 grouped convolution to restore the spatial feature size from $H' \times W' \times C''$ to $H \times W \times C'$ (with C'' input channels and $C'S^2$ output channels), followed by a pixel shuffle operation to recover the spatial resolution to $H \times W$. $\Phi^{CR}(\cdot)$ performs feature channel restoration through a pointwise convolution applied to the input features, maintaining identical spatial resolution and channel count to the input feature F, thereby yielding the final feature $\bar{F}$.

3.4. Interactive Fusion Module

To address the challenge of ineffective collaborative fusion between features from the CNN and Transformer dual-branch architectures, this paper proposes an Interactive Fusion Module whose structure is illustrated in Figure 5. The module adopts the channel attention mechanism and spatial attention mechanism from the SCConv method [27], establishing a multi-granularity feature interaction framework through the parallel deployment of a Channel Reconstruction Unit (CRU) and a Spatial Reconstruction Unit (SRU). The feature maps produced by the two sub-modules are concatenated, yielding the final feature map refined through both spatial and channel-wise operations. The attention maps generated by two submodules undergo dynamic coupling through a gated mechanism, achieving adaptive recalibration between the CNN branch's multi-level local features and the Transformer branch's global semantic characteristics. This dual-path attention synergy mechanism not only enhances local details within CNN-generated feature maps but also effectively suppresses noise, thereby establishing a geometry-preserving representation space that facilitates multi-scale feature fusion.

While the feature maps generated by the CNN branch contain rich local details, they may also exhibit noise. Therefore, the Channel Reconstruction Unit (CRU) is employed to meticulously refine these CNN-generated feature maps. The input spatially refined features X_w are partitioned into two components: Channel αC and Channel $(1-\alpha)C$, where α set experimentally to 0.5 serves as the hyperparameter. A 1×1 convolutional kernel is employed on each sub-path to perform sparse compression of feature channels, yielding X_{up} and X_{low}, respectively. Subsequently, grouped convolution (GWC) and pointwise convolution (PWC) are applied in parallel on Channel αC to achieve efficient feature recombination. The GWC effectively reduces parameter count and computational complexity, while the PWC compensates for potential information loss induced by grouped convolutions. This dual mechanism promotes information interaction and flow across feature channels, with the final output Y_1 obtained by summing the combined operations described above. Simultaneously, feature X_{low} from Channel $(1-\alpha)C$ is processed through PWC to generate supplementary features. The optimized features from the PWC output are

then fused with the original input features via channel-wise joint representation along the channel dimension, yielding an enhanced representation Y_2 through cross-source information fusion. Subsequently, spatial contextual information and channel-wise statistical characteristics are captured through global average pooling (GAP), yielding pooled results S_1 and S_2. These results are then subjected to a Softmax operation to derive normalized weight distributions, i.e., feature weight vectors β_1 and β_2, which are utilized to generate the final output $Y = \beta_1 Y_1 + \beta_2 Y_2$ via weighted feature aggregation. Finally, the weighted feature maps undergo secondary concatenation to generate a spatially correlated feature representation, thereby achieving the refined enhancement of localized attention weights.

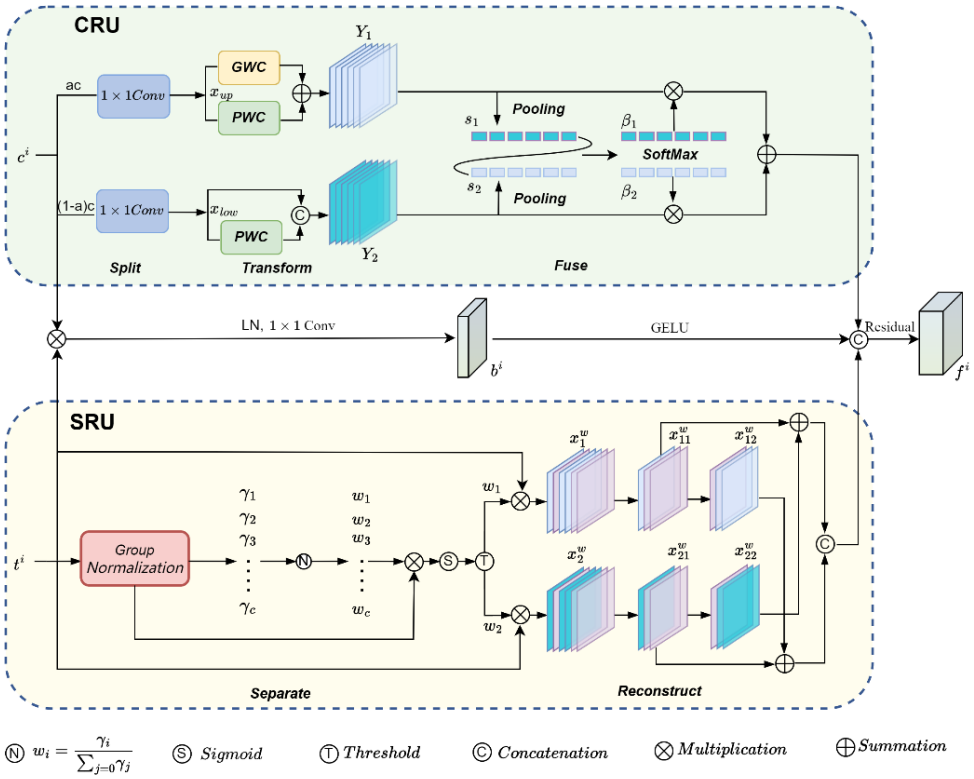


Figure 5. Interactive Fusion Module.

The self-attention mechanism in the Transformer architecture endows the Transformer branch with robust global contextual information capture capability, enabling its generated feature maps to effectively encompass rich global semantic information. To further enhance the integration and utilization of global information, this module introduces the SRU from the SCConv method to process the feature maps produced by the Transformer branch. By enhancing inter-channel correlations, the expression of global features is further optimized, thereby improving the model's capability to comprehend and process complex scenes with long-range dependencies. Initially, Group Normalization (GN) is applied to the input feature maps to differentiate between information-rich feature maps and information-sparse feature maps, as formally described by the following equation:

$$X_{\text{out}} = GN(X) = \gamma \frac{X - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta, \tag{10}$$

where $X \in \mathbb{R}^{N \times C \times H \times W}$ denotes the input feature map, where N is the batch size, C is the number of channels, and H and W are the spatial height and width. Respectively, μ and σ represent the group-wise mean and standard deviation, while ϵ is a small positive constant added to ensure numerical stability during the division operation. γ and β are trainable

affine transformation parameters. The scaling factor γ characterizes the spatial information richness of feature maps by quantifying the variance intensity of spatial pixel values across channels within a batch. Building on this, the weight coefficient W_γ is derived through normalization of γ, as defined in Equation (11):

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j}, \quad i, j = 1, 2, \ldots, C. \tag{11}$$

Subsequently, the Sigmoid function is applied to map the associated weight W_γ to the range $[0, 1]$, as defined in Equation (12):

$$W_\gamma = \text{Gate}(\text{Sigmoid}(W_\gamma(GN(X)))). \tag{12}$$

In subsequent processing, binary gating selection is performed by setting a discrimination threshold (experimentally set to 0.5). Based on this threshold, exceeding channels are fully activated ($W_1 = 1$) to retain informational weights, while under-activated channels are suppressed via zeroing ($W_2 = 0$) to form non-informational weights. Subsequently, the input feature X is multiplied by two distinct weighted features, yielding information-rich key features X_1^w and information-sparse redundant features X_2^w. To fully integrate the features from two weighted information streams, a cross-reconstruction mechanism is employed to achieve deep fusion of dual-channel weighted features. By enhancing the bidirectional interaction between key features and redundant features, the approach ultimately generates a spatially refined feature map X^w. Finally, the obtained Interactive Fusion Module b^i and branch features are concatenated and processed through a residual MLP to generate f^i.

The aforementioned operations partially mitigate common issues in traditional neural networks, such as gradient vanishing, gradient explosion, and network degradation. By effectively capturing multiple hierarchical levels and global and local feature information, these operations prevent the loss of edge-related information while ensuring the integrity and accuracy of feature representations.

4. Experimental Analysis

4.1. Dataset

This study employs the high quality public benchmark dataset from Zhang et al. [28]. As indicated in Table 1 and Figure 6, the dataset has 10 categories, consisting of 1500 panoramic dental X-ray images and their corresponding ground-truth segmentation masks, with each image measuring 1191 × 1211 pixels. Directly processing raw dental panoramic X-ray data through neural networks presents two challenges: (1) training with ultra-high-resolution images imposes significant computational burdens on hardware-constrained devices, frequently leading to GPU memory overflow or system RAM exhaustion and (2) when inputting images with unequal lengths and widths, splitting them first may disrupt important pixel relationships in the original images. To address these issues, all images are resized to 512 × 512 pixels. However, when resizing panoramic images to 512 × 512 pixels, the subtle inter-tooth spacing in original images may become compressed due to scaling, consequently increasing the model's difficulty in distinguishing boundaries. To mitigate this issue, the training procedure applies random horizontal flips, vertical flips, and their combined transformations to input data for enhancing the model's generalization capability and improving its adaptation to dental morphological variations. Furthermore, the designed dual-stream network utilizes the Transformer branch to capture spatial priors of tooth arrangement and the CNN branch to enhance local detail extraction. This dual-stream collaboration enables the network to preserve critical boundary features when

processing resized images, effectively mitigating the negative impact of image scaling on segmentation accuracy. Subsequently, the dataset was randomly partitioned into a training set (1200 images), a validation set (150 images), and a test set (150 images).

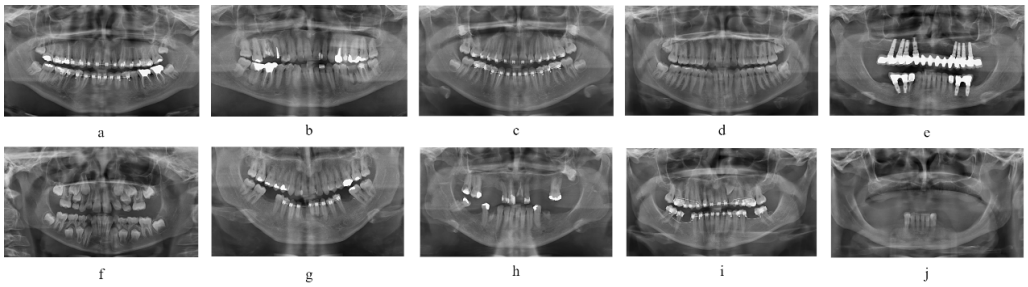


Figure 6. Panoramic X-ray dataset with 10 categories.

Table 1. Dental image classification.

Class	Tooth Type	Count
a	Restored normal dentition images with orthodontic appliances	73
b	Restored normal dentition images (excluding orthodontic appliances)	220
c	Normal dentition images with orthodontic appliances (excluding dental restorations)	45
d	Normal dentition images without dental restorations or orthodontic appliances	140
e	Dental images with dental implants	120
f	Dental images with supernumerary teeth (>32 teeth)	170
g	Edentulous images with dental restorations and orthodontic appliances	115
h	Edentulous images with dental restorations (excluding orthodontic appliances)	457
i	Edentulous images with orthodontic appliances (excluding dental restorations)	45
j	Edentulous images without dental restorations or orthodontic appliances	115

4.2. Experimental Setup

To quantitatively evaluate the performance of the proposed algorithm, four canonical evaluation metrics were systematically employed: Intersection over Union (*IoU*), *Precision*, *Accuracy*, and Dice coefficient (*Dice*). Among these metrics, the Dice coefficient is widely used in assessing medical image segmentation model performance. The computation of these metrics not only considers the overall segmentation performance but also calculates the mean values for the background and foreground separately, thereby providing a more comprehensive evaluation of the model's performance across different categories. The mathematical definitions of these four metrics are presented as follows:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{13}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{14}$$

$$Precision = \frac{TP}{TP + FP}, \tag{15}$$

$$Dice = \frac{TP + TP}{TP + TP + FP + FN}. \tag{16}$$

The definitions are as follows: TP (True Positive) denotes the number of pixels where both the prediction and the ground truth are dental regions; FP (False Positive) represents the number of pixels predicted as dental regions but actually belonging to the background; TN (True Negative) refers to the number of pixels where both the prediction and the ground truth are background regions; FN (False Negative) indicates the number of pixels predicted

as background regions but actually corresponding to dental regions. IoU is utilized to compute the ratio of overlap between the predicted segmentation results and ground-truth segmentation masks. Precision quantifies the proportion of pixels correctly identified as dental structures within all predicted tooth regions. Accuracy measures the percentage of correctly classified pixels (both true positives and true negatives) relative to the total number of pixels in the image. The Dice quantifies the similarity between the predicted segmentation results and ground-truth segmentation masks.

The experimental hardware environment is as follows: the CPU is an Intel Xeon Silver 4214R(Manufacturer: Intel Corporation; Santa Clara, USA; accessed via AutoDL, Shanghai, China); the GPU is an NVIDIA GeForce RTX 3080 Ti(Manufacturer: NVIDIA Corporation; Santa Clara, USA; accessed via AutoDL, Shanghai, China); and the system is equipped with 32 GB of RAM. The software environment includes a Windows 10 64-bit operating system, Python 3.8, PyTorch 1.8.0, and CUDA 11.6. The experiments can be executed on low-end graphics cards under identical software environments. The proposed method is implemented within the mmsegmentation codebase based on the PyTorch framework, where ViT-B16 is pretrained on the ImageNet-1K dataset, while the decoder component is randomly initialized.

In the weight initialization part, the Kaiming initialization method [29] is adopted to maintain the weight variance within a specific range, ensuring data stability and effectively controlling the values during both forward propagation and backward propagation, as shown in Equation (17). Here, the bias terms are initialized to 0, where $\hat{n}_i$ denotes the number of input parameters and W represents the weights,

$$W \sim \mathrm{N}\left[0, \sqrt{\frac{2}{\hat{n}_i}}\right]. \tag{17}$$

The network training settings are as follows: As shown in Equation (18), a dynamic learning rate adjustment strategy is adopted. The number of training epochs is set to 200, the Adam optimizer is employed for training the neural network, the batch size is set to 4, and the initial learning rate is set to 0.001. The current iteration count (iter) and the maximum iteration count (max_iter) are set to the current iteration and 40K, respectively, Additionally, the momentum (power) is set to 0.9,

$$\text{learning_rate} = base_lr \times \left(1 - \frac{iter}{\max_iter}\right)^{power}. \tag{18}$$

To accelerate model convergence, the cross-entropy loss function is adopted; a smaller loss value indicates a reduced discrepancy between predicted and ground truth values. As formalized in Equation (19), y denotes the model output,

$$L = -\mathrm{y}\log\mathrm{y} - (1-\mathrm{y})\log(1-\mathrm{y}). \tag{19}$$

4.3. Comparative Experiments

The experimental validation was conducted as follows: To comprehensively verify the advantages and effectiveness of the proposed algorithm, two sets of comparative experiments were implemented based on the aforementioned datasets and experimental configurations. The first set involved comprised experimental evaluations with U-Net-like architectures, while the second set comprised experimental evaluations against Transformer-based networks.

4.3.1. Comparative Experimental Analysis of U-Net-like Networks

To comprehensively evaluate and validate the advantages of the proposed method in the field of dental panoramic X-ray image segmentation, representative U-Net-type segmentation algorithms from recent years—including U-Net [16], U-Net++ [30], Attention U-Net [31], CE-Net [32], and GT U-Net [21]—were selected for comparative analysis. As a classical benchmark model in medical image segmentation, U-Net establishes a critical foundation for subsequent algorithmic improvements through its distinctive encoder–decoder architecture and skip connection mechanisms. Building upon this framework, U-Net++ introduces nested skip connections and a deep supervision mechanism, enhancing the model's representational capacity by incorporating a multi-scale hierarchical architecture. Attention U-Net integrates the dual mechanisms of channel attention and spatial attention into the original U-Net framework, dynamically adjusting feature weight distributions to significantly enhance the network's focus on critical features. In contrast, CE-Net adopts adaptive convolution methods for image segmentation, aiming to improve the network's understanding of complex contextual patterns. Notably, GT U-Net, an optimized solution specifically designed for root segmentation in panoramic X-ray images, demonstrated superior performance in dental root edge localization. Table 2 presents the comparative segmentation accuracy of different algorithms. Compared to the top-performing Attention U-Net model, the proposed method achieves improvements of 0.77% in IoU, 1.65% in Precision, 0.21% in Accuracy, and 0.67% in Dice coefficient. Compared with the experimental results of U-Net, CE-Net, GT U-Net, and U-Net++, the proposed method demonstrated improvements across four evaluation metrics: IoU, Precision, Accuracy, and Dice. These results empirically validate that, among U-Net-like methods, the proposed approach achieves superior comprehensive segmentation performance in dental image segmentation, exhibiting an enhanced capability to distinguish between dental target regions and background areas.

Table 2. Segmentation accuracy comparison of U-Net-like algorithms on dental panoramic X-ray images (Bold values indicate optimal results, same for the tables below).

Method	IoU (%)	Precision (%)	Accuracy (%)	Dice (%)	Params (M)	FPS
U-Net	85.73	93.45	90.89	92.10	**7.7**	**152.4**
GT UNet	85.70	91.66	95.74	92.56	32.1	81.2
CE Net	88.47	94.57	95.26	93.63	29.7	123.5
U-Net++	90.48	94.41	97.09	93.48	9.12	105.8
Att U-Net	90.79	93.05	97.22	93.91	8.75	120.0
Ours	**91.49**	**94.59**	**97.42**	**94.54**	17.4	136.6

To intuitively demonstrate the segmentation performance across different algorithms, the segmentation results of various methods on the test set were visualized. Figure 7 illustrates the overall segmentation outcomes of U-Net-like algorithms, while Figure 8 provides a detailed visualization of their localized segmentation performance. As observed in the figures, U-Net exhibits significant intra-layer semantic feature discrepancies and is constrained by insufficient feature extraction capability, leading to fragmented segmentation outcomes when delineating complete dental structures. The visualization results of U-Net++ exhibited noticeable isolated segmentation errors, primarily attributed to insufficient multi-scale information extraction. Both Attention U-Net and CE-Net demonstrated pronounced serrated edges in dental segmentation boundaries and suffered from over-segmentation issues. Isolated segmentation errors and partial dental adhesion artifacts were observed in the visualization results of GT U-Net, which were attributed to the partial

loss of critical receptive field information in the network architecture. Compared with other methods, the proposed method produces dental structures closer to the ground-truth labels. While preserving precise dental boundaries and complete anatomical integrity, the method achieves the following refinements: (1) smoothed edge continuity with alleviated serrated artifacts, (2) mitigated over-segmentation tendencies, and (3) effective suppression of isolated segmentation errors.

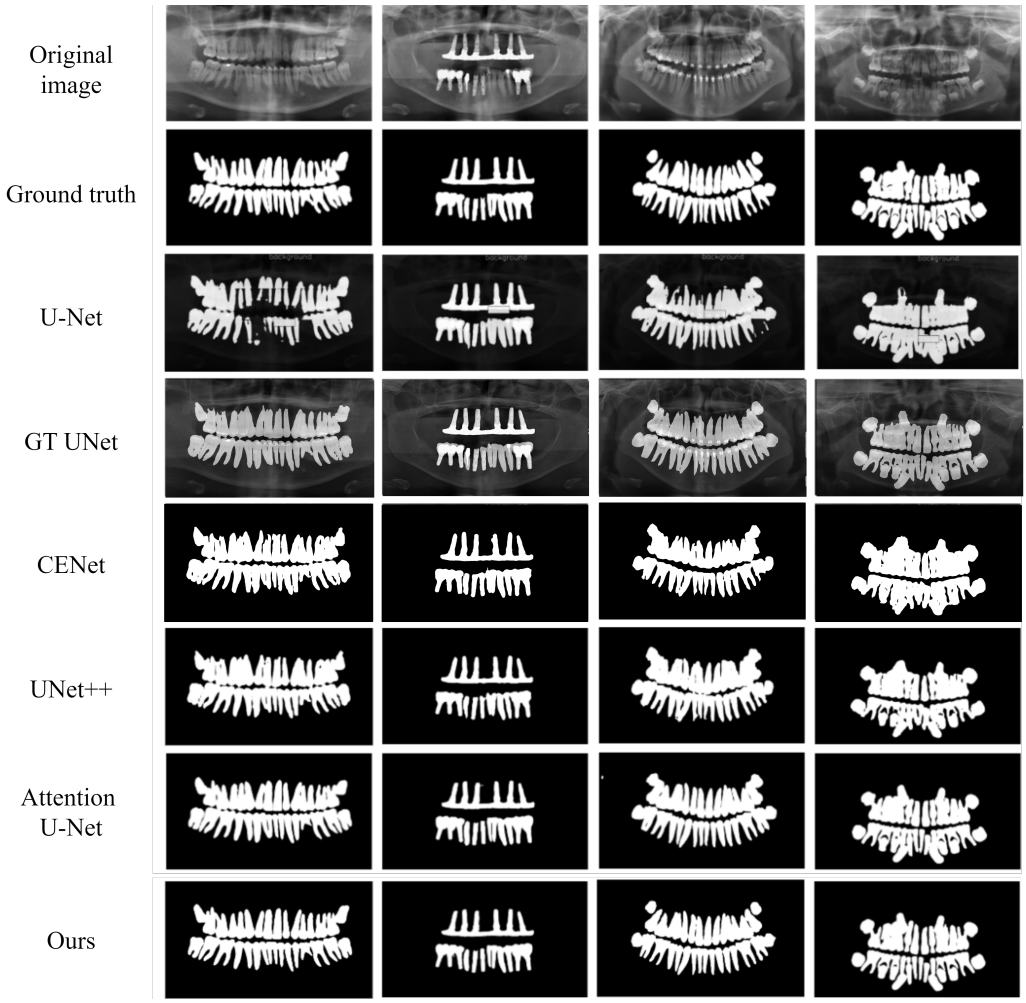


Figure 7. Holistic segmentation results visualization of U-Net-like algorithms.

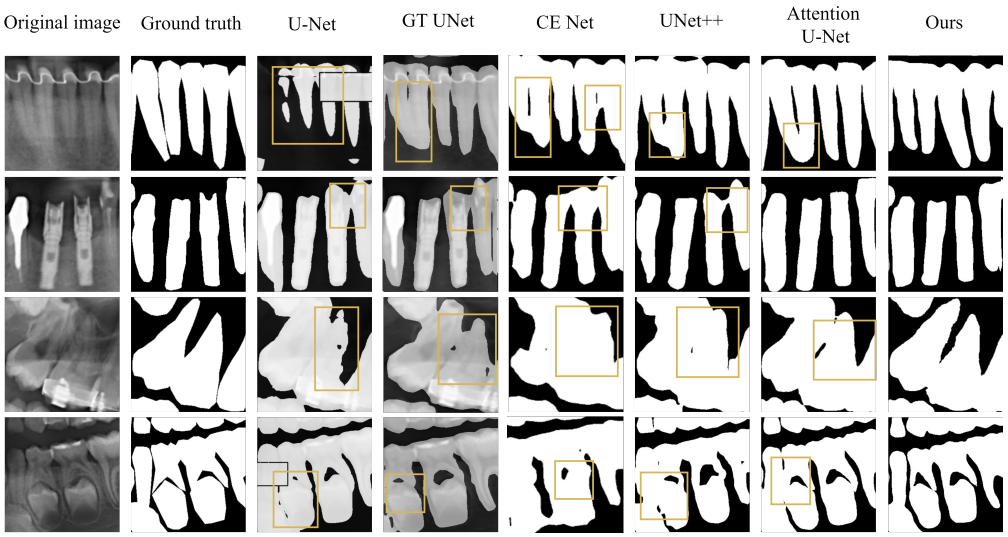


Figure 8. Visualization of fine-grained segmentation results from the U-Net-based algorithm.

4.3.2. Comparative Experimental Analysis of Transformer-like Networks

Having validated the superiority of the proposed method in U-Net-like networks, we further investigated its performance in Transformer-based architectures by conducting comparative analyses with Polyer [33] and state-of-the-art Transformer-integrated segmentation algorithms, including Swin-T [34], TopFormer [35], SegNext [36], Afformer [37], and SeaFormer [38]. Table 3 presents the comparative results of the segmentation accuracy of dental panoramic X-ray images among the different algorithms. The algorithm achieves optimal performance across all evaluation metrics, including IoU, Dice coefficient, Precision, and Accuracy. Compared with Swin-T, which demonstrates the best segmentation performance among the compared methods, the approach achieves improvements of 1.13%, 1.43%, 2.27%, and 0.97% in IoU, Precision, Accuracy, and Dice, respectively.

Table 3. Comparison of segmentation accuracy of Transformer-based algorithms on dental panoramic radiographs.

Method	IoU (%)	Precision (%)	Accuracy (%)	Dice (%)	Params (M)	FPS
Polyer	87.10	90.66	95.94	92.96	28.0	138.9
Swin-T	90.47	93.57	95.26	93.63	28.3	128.4
TopFormer	88.19	93.14	94.06	93.59	5.7	106.7
SegNext	87.77	92.89	93.80	93.34	27.7	78.3
Afformer	88.03	93.51	93.47	93.49	3.0	**148.4**
SeaFormer	85.70	91.92	92.27	92.09	**1.8**	103.6
Ours	**91.49**	**94.91**	**97.42**	**94.54**	17.4	136.6

The segmentation results are visualized for diverse image types in the test set, including adult healthy teeth, dental implants, and missing teeth. Figure 9 presents the global segmentation visualizations of the Transformer-like algorithm, while Figure 10 illustrates its detailed segmentation performance in localized regions. The visual analysis reveals notable deficiencies in Polyer's performance regarding dental segmentation tasks. The visualizations demonstrate significant limitations in Polyer's performance for dental segmentation tasks, where substantial inter-layer semantic feature discrepancies result in critical information loss during the processing of complex anatomical structures. The segmentation outputs of Swin-T exhibit deficiencies in multi-scale feature extraction, leading to inaccuracies in detail preservation that often manifest as isolated missegmentation artifacts within complex backgrounds or boundary-proximal regions. The dental segmentation results of TopFormer exhibit noticeable edge jaggedness artifacts and a tendency toward over-segmentation. Meanwhile, the multi-scale convolutional attention module in SegNext introduces heightened architectural complexity and an expanded parameter count, resulting in substantial computational overhead during both training and inference phases, as well as insufficient generalization capability. These limitations collectively degrade segmentation accuracy. The Afformer network introduces adaptive frequency filters into segmentation tasks, theoretically enabling enhanced capture of frequency information within images to improve segmentation accuracy. However, in practical applications, imbalanced handling of low- and high-frequency information may may cause insufficiently refined edge delineation and partial under-segmentation in the resulting outputs. Compared to other algorithms, the proposed method demonstrates distinct advantages in visualized segmentation outcomes. Its reconstructed dental geometric topology aligns more closely with ground-truth annotations, accurately preserving tooth boundaries and structural integrity. The geometric topological structure of the teeth demonstrates closer alignment with ground-truth labels, precisely preserving dental boundaries and complete anatomical configurations. This methodology

achieves enhanced clarity in tooth root boundaries, eliminates visible adhesion between adjacent teeth, and reduces jagged edge artifacts along dental margins.

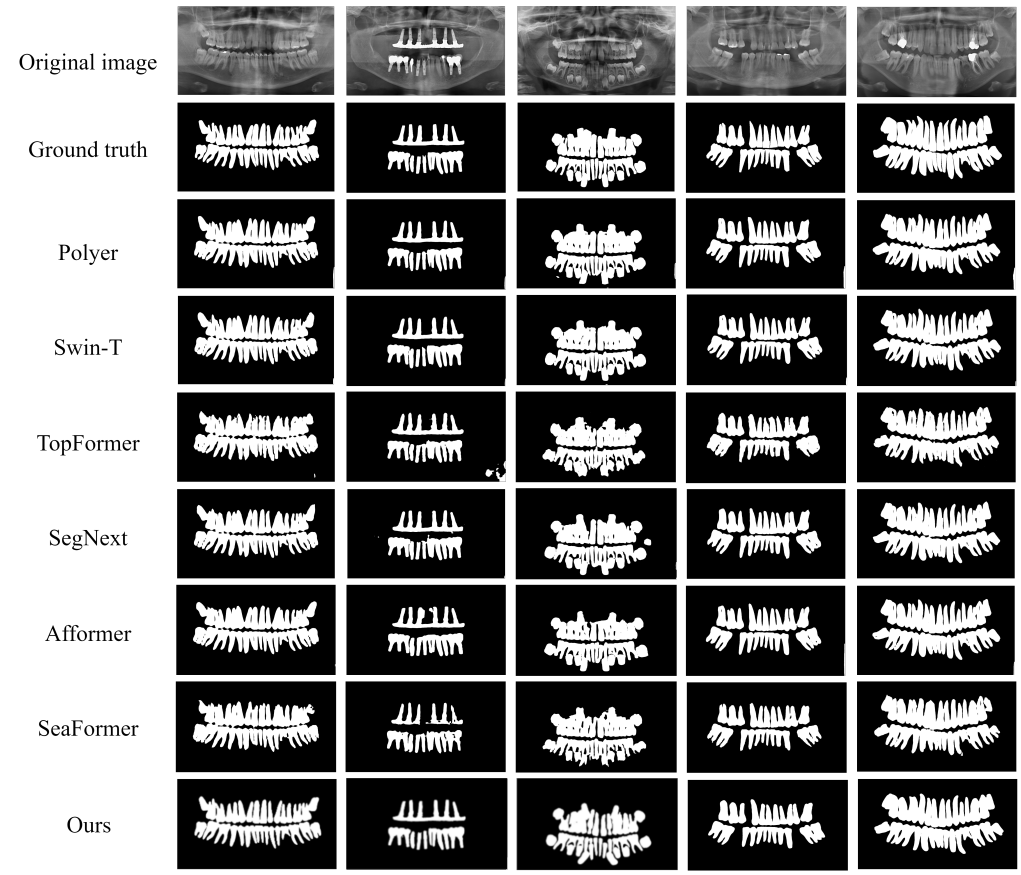


Figure 9. Visualization of macroscopic segmentation results from the Transformer-based algorithm.

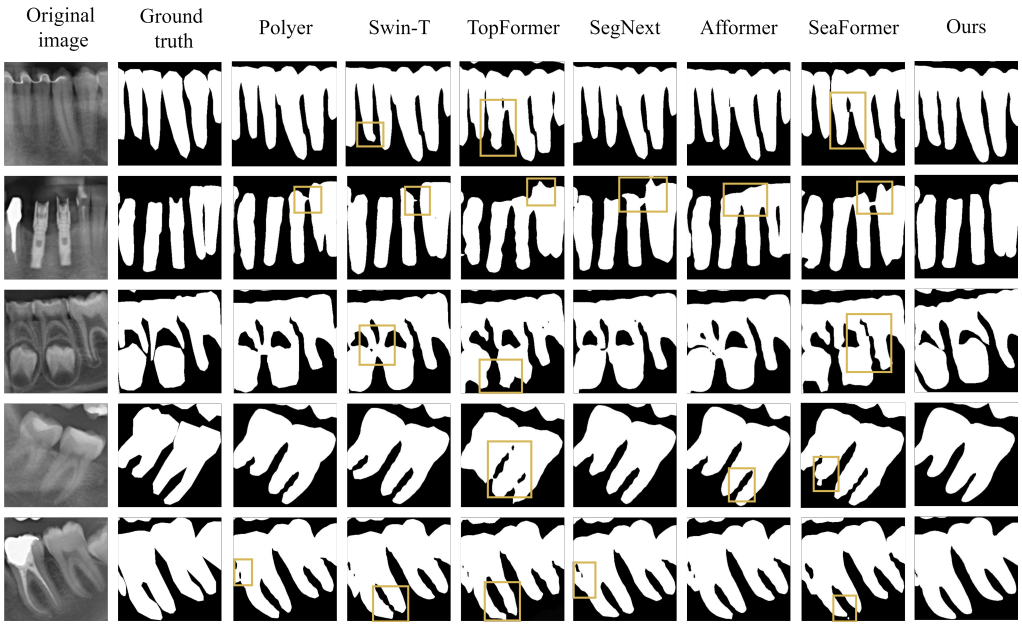


Figure 10. Visualization of fine-grained segmentation results from the Transformer-based algorithm.

4.4. Ablation Study Analysis

4.4.1. Ablation Study on the Pooling-Cooperative Convolutional Module

The pooling collaborative convolution module innovatively incorporates median pooling operation alongside conventional global average pooling and global max pooling.

Through computation of channel-wise attention weights, this module achieves intensified concentration on localized contextual semantics, thereby achieving the refined capture of intrinsic correlations among dental edge features across different channels. Experimental results demonstrate that the designed pooling collaborative convolution module achieves improvements of 1.86 percentage points and 1.32 percentage points in the IoU and Dice coefficient metrics, respectively. To thoroughly investigate the specific effects of the three pooling operations, we perform ablation studies on each individual pooling component within the pooling collaborative convolution module, with the experimental results presented in the table below.

As evidenced by the experimental data in Table 4, the Pooling-Cooperative Convolutional Module attains optimal comprehensive performance when retaining the complete set of max pooling, average pooling, and median pooling operations. Experimental results indicate that compared to configurations retaining only max pooling and average pooling operations, the proposed module demonstrates improvements of 0.43 and 0.58 percentage points in the IoU and Dice coefficients, respectively. This evidence suggests that the median pooling operation effectively leverages its unique statistical properties to extract median-value features inaccessible to conventional pooling methods, particularly exhibiting superior performance in dental edge noise suppression and local detail preservation. Furthermore, when the module exclusively employs a single pooling operation, the CNN branch exhibits constrained representational capacity for textural features and geometric structural information at dental edges, owing to the absence of a multi-scale feature complementarity mechanism.

Table 4. Ablation study on the Pooling-Cooperative Convolutional Module.

Max Pooling	Average Pooling	Median Pooling	IoU (%)	Dice (%)
			89.08	92.81
✓			90.59	93.37
	✓		89.85	92.79
		✓	89.41	92.06
	✓	✓	89.66	93.16
✓		✓	89.51	93.06
✓	✓		90.31	93.46
✓	✓	✓	**90.74**	**94.04**

4.4.2. Ablation Study on the Semantic Transformation Module

To address the challenge of Transformer architectures ineffectively capturing global dependencies within shallow network layers, the designed Semantic Transformation Module is deployed during deep feature extraction stages. To systematically validate the effectiveness of the proposed module, multiple controlled ablation experiments were designed for real-time segmentation tasks. By substituting the Semantic Transformation Module with five distinct types of feature extraction modules (including diverse convolutional blocks and Transformer variants), comprehensive performance comparisons were conducted, with the results detailed in Table 5. The proposed Semantic Transformation Module achieves superior segmentation accuracy compared to the classical SegFormer [39] and demonstrates enhanced precision over existing methods including CFBlock [40], GFABlock [41], and MSCANBlock [36]. The experimental results demonstrated that the designed Semantic Transformation Module achieves state-of-the-art performance across the IoU, Precision, Accuracy, and Dice metrics, substantiating its effectiveness in mitigating the semantic information gap between CNN and Transformer architectures.

Table 5. Accuracy comparison of different architectural modules.

Module	IoU (%)	Precision (%)	Accuracy (%)	Dice (%)
SegFormer	89.17	93.46	96.62	93.15
GFABlock	88.85	92.54	95.17	92.69
MSCANBlock	87.69	93.64	95.64	92.53
CFBlock	89.31	93.62	96.82	93.16
STM (Ours)	**89.79**	**93.91**	**97.03**	**93.54**

4.4.3. Ablation Study on the Interactive Fusion Module

To investigate the critical role of the designed Interactive Fusion Module in enhancing dental medical image segmentation performance, hierarchical ablation experiments were conducted on the dataset. The segmentation results of the baseline model, the model with only the channel attention mechanism (CRU) added, the model with only the spatial attention mechanism (SRU) added, and the model with the Interactive Fusion Module added were compared, respectively.

The statistical results in Table 6 indicate that removing the Interactive Fusion Module (as the baseline model) and directly concatenating features from the two branches leads to reductions of 2.57 percentage points in IoU and 0.71 percentage points in the Dice coefficient, thereby intuitively confirming the critical importance of the Interactive Fusion Module in feature fusion processes. To conduct an in-depth analysis of the effects of different components in the Interactive Fusion Module on segmentation performance, a component ablation study was further implemented. When only the CRU was preserved, the model exhibited decreases of 2.26 percentage points in IoU and 0.42 percentage points in the Dice coefficient for medical image segmentation tasks. In contrast, when only the SRU was retained, reductions of 2.14 percentage points in IoU and 0.16 percentage points in the Dice coefficient were observed. This demonstrates that the integration of channel features and spatial information effectively enhances the segmentation performance of CNN-based and Transformer-based dual-stream networks, highlighting the superiority of the proposed method in dental medical image segmentation.

Table 6. Ablation analysis of the Interactive Fusion Module.

Method	IoU (%)	Precision (%)	Accuracy (%)	Dice (%)
w/o IFM	87.88	92.87	92.74	92.81
w/CRU	88.19	92.40	96.62	93.10
w/SRU	88.31	92.72	96.42	93.36
IFM (Ours)	**90.45**	**93.71**	**96.62**	**93.52**

5. Discussion

This paper proposes a Transformer-based dual-stream dental panoramic X-ray image segmentation network with heterogeneous feature complementarity that effectively enhances tooth localization capability and segmentation accuracy by constructing a global–local feature co-enhancement mechanism. This network adopts a dual-branch parallel architecture, where the CNN branch focuses on local texture feature extraction, while the Transformer branch employs the self-attention mechanism to capture global spatial dependencies. The designed Semantic Transformation Module and Interactive Fusion Module enable interaction between the two heterogeneous branches in multi-level feature spaces, which not only overcomes the limited local receptive fields of traditional CNN models but also compensates for the deficiencies of Transformer architectures in local detail modeling and translation invariance, thereby acquiring more comprehensive global information. The experimental results demonstrate that on the public dental image dataset benchmark

quantitative evaluation, the proposed method achieves 91.49% and 94.54% on the IoU and Dice metrics. The experimental results demonstrate that the proposed method achieves superior performance over existing state-of-the-art techniques in dental segmentation tasks, producing segmentation results with closer approximation to ground-truth annotations while attaining higher accuracy.

Although promising results have been achieved in fully supervised dental medical image segmentation tasks, the number of parameters has not yet reached the optimal level of lightweight models, and validation has not been conducted on datasets with diverse imaging modalities. Future work will focus on model compression strategies and multi-source data generalization. Through approaches such as lightweight architecture replacement and knowledge distillation optimization, the computational complexity is reduced while maintaining the existing segmentation accuracy, with the objective of achieving superior performance metrics across diverse datasets. Furthermore, the annotation of dental data remains a time-consuming and costly task, which limits the feasibility of processing large-scale datasets. Therefore, semi-supervised segmentation methods are of particular importance, as they can maintain segmentation performance while reducing the reliance on annotated data by utilizing partially annotated dental image data. Future work will focus on investigating self-training and pseudo-labeling techniques to enhance the accuracy of dental segmentation under semi-supervised learning frameworks.

Author Contributions: Conceptualization, T.M.; methodology, T.M. and J.L.; software, J.L. and Z.D.; validation, Y.L. (Yawen Li); writing—original draft preparation, Z.D.; writing—review and editing, T.M. and J.L.; supervision, Y.L. (Yuancheng Li); funding acquisition, T.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Innovation 2030—New Generation Artificial Intelligence Major Project (No. 2022ZD0119005, led by Lin Jiang), The Youth Innovation Team of Shaanxi Universities, and the Shaanxi Natural Science Fundamental Research Program Project (No. 2022JM-508, led by Tian Ma).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Figshare at https://doi.org/10.6084/m9.figshare.c.6317013.v1 (Access Date: 2024 November 20), reference number 6317013. These data were derived from the following resources available in the public domain: Zhang, Yifan; Ye, Fan; Chen, Lingxiao; et al. "Children's Dental Panoramic Radiographs Dataset for Caries Segmentation and Dental Disease Detection" (figshare Collection, 2023) [28].

Acknowledgments: The authors are grateful to the editor and reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lin, P.; Huang, P.; Huang, P.; Hsu, H.; Chen, C. Teeth segmentation of dental periapical radiographs based on local singularity analysis. *Comput. Methods Programs Biomed.* **2014**, *113*, 433–445. [CrossRef] [PubMed]
2. Park, K.J.; Kwak, K.C. A trends analysis of dental image processing. In Proceedings of the 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 20–22 November 2019; IEEE: Piscataway, NJ, USA, 2019: pp. 1–5.
3. Arora, S.; Tripathy, S.K.; Gupta, R.; Srivastava, R. Exploiting multimodal CNN architecture for automated teeth segmentation on dental panoramic X-ray images. *Proc. Inst. Mech. Eng. Part H J. Eng. Med.* **2023**, *237*, 395–405. [CrossRef] [PubMed]
4. Hou, S.; Zhou, T.; Liu, Y.; Dang, P.; Lu, H.; Shi, H. Teeth U-Net: A segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement. *Comput. Biol. Med.* **2023,** *152,* 106296. [CrossRef] [PubMed]
5. Chandrashekar, G.; AlQarni, S.; Bumann, E.E.; Lee, Y. Collaborative deep learning model for tooth segmentation and identification using panoramic radiographs. *Comput. Biol. Med.* **2022**, *148*, 105829. [CrossRef] [PubMed]

6. Li, P.; Gao, C.; Lian, C.; Meng, D. Spatial Prior-Guided Bi-Directional Cross-Attention Transformers for Tooth Instance Segmentation. *IEEE Trans. Med. Imaging* **2024**, *43*, 3936–3948. [CrossRef] [PubMed]
7. Ma, T.; Yang, Y.; Zhai, J.; Yang, J.; Zhang, J. A tooth segmentation method based on multiple geometric feature learning. *Healthcare* **2022**, *10*, 2089. [CrossRef] [PubMed]
8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
9. Radhiyah, A.; Harsono, T.; Sigit, R. Comparison study of Gaussian and histogram equalization filter on dental radiograph segmentation for labelling dental radiograph. In Proceedings of the 2016 International Conference on Knowledge Creation and Intelligent Computing (KCIC), Manado, Indonesia, 15–17 November 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 253–258.
10. Tikhe, S.V.; Naik, A.M.; Bhide, S.D.; Saravanan, T.; Kaliyamurthie, K. Algorithm to identify enamel caries and interproximal caries using dental digital radiographs. In Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, 27–28 February 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 225–228.
11. Alsmadi, M.K. A hybrid Fuzzy C-Means and Neutrosophic for jaw lesions segmentation. *Ain Shams Eng. J.* **2018**, *9*, 697–706. [CrossRef]
12. Trivedi, D.N.; Modi, C.K. Dental contour extraction using ISEF Algorithm for human identification. In Proceedings of the 2011 3rd International Conference on Electronics Computer Technology, Kanyakumari, India, 8–10 April 2011; IEEE: Piscataway, NJ, USA, 2011; Volume 6, pp. 6–10.
13. Modi, C.K.; Desai, N.P. A simple and novel algorithm for automatic selection of ROI for dental radiograph segmentation. In Proceedings of the 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE), Niagara Falls, ON, Canada, 8–11 May 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 000504–000507.
14. Haghanifar, A.; Majdabadi, M.M.; Haghanifar, S.; Choi, Y.; Ko, S.B. PaXNet: Tooth segmentation and dental caries detection in panoramic X-ray using ensemble transfer learning and capsule classifier. *Multimed. Tools Appl.* **2023**, *82*, 27659–27679. [CrossRef]
15. Tekin, B.Y.; Ozcan, C.; Pekince, A.; Yasa, Y. An enhanced tooth segmentation and numbering according to FDI notation in bitewing radiographs. *Comput. Biol. Med.* **2022**, *146*, 105547.
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer: Cham, Switzerland, 2015; pp. 234–241.
17. Koch, T.L.; Perslev, M.; Igel, C.; Brandt, S.S. Accurate segmentation of dental panoramic radiographs with U-Nets. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 15–19.
18. Zhao, Y.; Li, P.; Gao, C.; Liu, Y.; Chen, Q.; Yang, F.; Meng, D. TSASNet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network. *Knowl.-Based Syst.* **2020**, *206*, 106338. [CrossRef]
19. Ma, T.; Zhou, X.; Yang, J.; Meng, B.; Qian, J.; Zhang, J.; Ge, G. Dental lesion segmentation using an improved icnet network with attention. *Micromachines* **2022**, *13*, 1920. [CrossRef] [PubMed]
20. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
21. Li, Y.; Wang, S.; Wang, J.; Zeng, G.; Liu, W.; Zhang, Q.; Jin, Q.; Wang, Y. Gt u-net: A u-net like group transformer network for tooth root segmentation. In Proceedings of the Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, 27 September 2021; Proceedings 12; Springer: Cham, Switzerland, 2021; pp. 386–395.
22. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. Unetr: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
23. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision 2022, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 205–218.
24. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
25. Ma, T.; Dang, Z.; Yang, Y.; Yang, J.; Li, J. Dental panoramic X-ray image segmentation for multi-feature coordinate position learning. *Digit. Health* **2024**, *10*, 20552076241277154. [CrossRef] [PubMed]
26. Zhao, H.; Gou, Y.; Li, B.; Peng, D.; Lv, J.; Peng, X. Comprehensive and delicate: An efficient transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 14122–14132.

27. Li, J.; Wen, Y.; He, L. Scconv: Spatial and channel reconstruction convolution for feature redundancy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 6153–6162.
28. Zhang, Y.; Ye, F.; Chen, L.; Xu, F.; Chen, X.; Wu, H.; Cao, M.; Li, Y.; Wang, Y.; Huang, X. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. *Sci. Data* **2023**, *10*, 380. [CrossRef] [PubMed]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
30. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4*; Springer: Cham, Switzerland, 2018; pp. 3–11.
31. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
32. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
33. Shao, H.; Zhang, Y.; Hou, Q. Polyper: Boundary sensitive polyp segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence 2024, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4731–4739.
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
35. Zhang, W.; Huang, Z.; Luo, G.; Chen, T.; Wang, X.; Liu, W.; Yu, G.; Shen, C. Topformer: Token pyramid transformer for mobile semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 12083–12093.
36. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022,** *35*, 1140–1156.
37. Dong, B.; Wang, P.; Wang, F. Head-free lightweight semantic segmentation with linear transformer. In Proceedings of the AAAI Conference on Artificial Intelligence 2023, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 516–524.
38. Wan, Q.; Huang, Z.; Lu, J.; Yu, G.; Zhang, L. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
39. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
40. Xu, Z.; Wu, D.; Yu, C.; Chu, X.; Sang, N.; Gao, C. Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence 2024, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6378–6386.
41. Wang, J.; Gou, C.; Wu, Q.; Feng, H.; Han, J.; Ding, E.; Wang, J. Rtformer: Efficient design for real-time semantic segmentation with transformer. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 7423–7436.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.