

Article

DoubleTransU-Net: Enhancing Teeth Segmentation in Panoramic Dental X-Ray Images

Manal Touahri * and Aissam Berrahou

ENSIAS, Mohammed V University, Rabat 10112, Morocco

* Correspondence: manal_touahri@um5.ac.ma

Abstract

Accurate teeth segmentation in panoramic dental radiographs remains a challenging task due to high image noise, low contrast, the similarity in intensity between teeth and surrounding tissues, and blurred tooth boundaries. To address these challenges, we propose DoubleTransU-Net, a dual-stage hybrid CNN–Transformer architecture that combines progressive segmentation refinement with global contextual feature learning. The first stage generates an initial tooth segmentation, while the second stage progressively refines ambiguous tooth regions to improve boundary delineation and segmentation accuracy. In addition, Atrous Spatial Pyramid Pooling (ASPP) modules capture multi-scale contextual information, whereas squeeze-and-excitation (SE) blocks enhance discriminative feature representations through channel-wise feature recalibration. The proposed model was evaluated on two public panoramic dental X-ray datasets, UFBA-UESC (1500 images) and Tufts (1000 images), and compared against several state-of-the-art segmentation models, including U-Net, DoubleU-Net, Attention U-Net, TransUNet, and DeepLabv3+. On the UFBA-UESC dataset, DoubleTransU-Net achieved an Accuracy of 95.54%, a Dice coefficient of 93.72%, an Intersection over Union (IoU) of 88.18%, a Precision of 93.57%, and a Recall of 94.24%. On the Tufts dataset, it achieved an Accuracy of 91.83%, a Dice coefficient of 92.93%, an IoU of 86.80%, a Precision of 92.27%, and a Recall of 93.97%. These results demonstrate that DoubleTransU-Net consistently outperforms existing state-of-the-art segmentation methods while exhibiting strong robustness and generalization across different panoramic dental datasets, highlighting its effectiveness for tooth semantic segmentation in panoramic dental X-ray images.

Keywords: tooth segmentation; dental panoramic radiographs; convolutional neural networks; transformers; public datasets



Academic Editor: Maryam Ravan

Received: 5 July 2026

Revised: 5 August 2026

Accepted: 9 August 2026

Published: 15 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Tooth segmentation in radiographic images has emerged as a key focus considering the widespread demand for solutions applied in multiple clinical applications, such as age and sex identification, orthodontic treatment, disease diagnosis and treatment planning. However, this process requires significant time and depends largely on the dentist's expertise and knowledge. In this regard, accurate tooth segmentation in dental images can greatly improve and accelerate the efficiency flow of clinical work by assisting in clinical diagnosis, help developing appropriate surgical and treatment plans and significantly reducing the workload of dentists.

Panoramic dental radiography is now a commonly utilized diagnostic tool in dental practice owing to its low radiation exposure and high cost-effectiveness. It serves an

essential role in aiding dentists to evaluate the gums, teeth, jaws and surrounding bone structure. Nonetheless, interpreting panoramic dental radiographs continues to pose significant challenges due to issues such as: (1) high noise and low contrast due to radiation limitation, (2) variations of teeth in different patients and dental instruments used for prostheses and restorations, and (3) the homogeneity of intensity in regions near teeth, where teeth have similar intensity to surrounding tissues and alveolar bone, which leads to teeth adherence and unclear boundaries in tooth prediction.

Recent developments in deep learning have greatly improved medical image segmentation performance, giving rise to some deep learning-based tooth segmentation methods particularly in semantic and instance-level tasks. Several works have employed the UFBA-UESC dental dataset [1] for teeth segmentation, addressing both semantic segmentation [2–4] and instance segmentation [5–7]. Additional approaches have explored advanced methodologies, including pixel-level semantic segmentation [8] and transformer-based models [9], to address the challenges of teeth segmentation in panoramic dental X-rays. Despite these advances, most existing studies evaluate their methods on a single dataset, which limits the assessment of their cross-dataset generalization. In addition, CNN-based approaches may have difficulty capturing long-range contextual information, whereas existing Transformer-based methods remain relatively underexplored for tooth segmentation in panoramic radiographs. To address these limitations, this study proposes DoubleTransU-Net, a dual-stage CNN–Transformer architecture for semantic tooth segmentation, and evaluates it on the UFBA-UESC and Tufts dental datasets.

In summary, the main contributions of this work are as follows:

- We propose DoubleTransU-Net, a dual-stage CNN–Transformer architecture that combines progressive mask refinement, global contextual modeling, multi-scale feature extraction through ASPP, and channel-wise attention through squeeze-and-excitation blocks.
- We conduct a comprehensive evaluation on two publicly available panoramic dental datasets and compare the proposed architecture with U-Net, DoubleU-Net, Attention U-Net, TransUNet, and DeepLabv3+ using Accuracy, Dice coefficient, IoU, Precision, and Recall.
- We investigate cross-dataset generalization by training the compared models on the Tufts dataset and evaluating them on unseen UFBA-UESC images, providing an assessment of their ability to handle differences between data distributions.

The remainder of this paper is organized as follows: In Section 2, we review related works. In Section 3, we present the materials and methods, detailing the datasets used, the proposed network architecture, the loss functions used and the implementation details. Section 4 discusses our results, including comparative performance across datasets, an ablation study, generalization experiments, and model complexity. This is followed by the discussion in Section 5, where we analyze the findings and their implications. Finally, the conclusion is drawn in Section 6.

2. Related Work

Recent progress in deep learning has considerably advanced medical image segmentation performance, leading to several deep learning-based tooth segmentation methods, particularly in semantic and instance-level tasks. Several studies have explored semantic segmentation of teeth utilizing the UFBA-UESC dental dataset [1]. Koch et al. [2] developed a network by combining four U-Net models into an ensemble and divided the original images into patches to serve as network inputs. The results were further improved by using data augmentation and test-time augmentation, reaching a Dice score of 93.6%. Zhao et al. [3] proposed TSASNet, a novel method that employs a dual-phase attention-based

segmentation architecture which performs teeth localization as the first stage by using an attention model, then segments the tooth region from the attention maps obtained from the first stage by using a fully convolutional network [10], achieving a 92.72% Dice score. More recently, Chen et al. [4] introduced MSLPNet, a teeth segmentation architecture that incorporates multi-scale location perception to address challenges with fuzzy boundaries. The approach was further enhanced by a hybrid loss function and achieved a Dice score of 93.01%. While these methods perform well, they primarily focus on single datasets and lack robust generalization across diverse data distributions. Additionally, their reliance on CNNs limits their ability to capture global contextual information, which is crucial for addressing the challenges of panoramic X-rays.

Instance segmentation of teeth has also been explored and three studies used the UFBA-UESC dental dataset for this purpose. Jader et al. [5] initiated the first research focused on instance-level segmentation of teeth despite the absence of tooth numbering. Mask R-CNN [11] with a pretrained ResNet-101 backbone was trained on 193 images, validated using 83 images, then evaluated on the remaining 1224 images reaching 88% of F1-score. Silva et al. [6] attempted to number and segment teeth on panoramic radiographs. The performance of four architectures were compared: PANet [12], HTC [13], Mask R-CNN and ResNeSt [14]. The PANet architecture provided the best results with an F1-score of 91.65%. To enhance the segmentation on the instance boundaries, Pinherio et al. [7] annotated a set of 450 radiographs from the UFBA-UESC dental dataset from scratch to account for overlapping instances and deciduous teeth. Two Mask R-CNN approaches were compared: the standard configuration with a fully convolutional network segmentation head, and an alternative utilizing the PointRend [15] module. The PointRend achieved 77.3% and 75.3% mAP scores on instance segmentation and the numbering tasks, respectively, outperforming the standard Mask R-CNN by margins of 1.2% and 2.0%. More recently, Rubiu et al. [16] also investigated using Mask R-CNN for teeth instance segmentation on the Tufts dental dataset [17] and attained a Dice score of 87%. Although these studies tackle instance segmentation, their approaches to data processing and modeling provide valuable insights relevant to semantic segmentation tasks.

Existing tooth-segmentation studies have investigated CNN ensembles, attention-based networks, generative adversarial networks, instance-segmentation frameworks, and Transformer-based architectures. However, the literature still presents several limitations. Many methods are evaluated on a single dataset, making their ability to generalize across different image distributions difficult to assess. CNN-based architectures are effective at extracting local spatial features but may have limited ability to model long-range contextual relationships, while existing Transformer-based approaches have shown variable performance on panoramic dental images. Furthermore, panoramic radiographs remain challenging because of image noise, low contrast, intensity similarity between teeth and surrounding tissues, and unclear tooth boundaries. These limitations motivate the development of methods that effectively combine local feature extraction with global contextual modeling while evaluating their robustness across multiple public datasets.

3. Materials and Methods

3.1. Datasets

3.1.1. Tufts Dental Dataset

The Tufts dental dataset [17] comprises 1000 fully anonymized and de-identified dental panoramic radiographs gathered between January 2014 and December 2016 from Tufts University. This multimodal dataset includes the teeth, abnormalities and maxillomandibular area outlined as well as the radiologist's expertise captured as gaze maps and audio-text files. Figure 1 illustrates example dataset images with their respective segmentation masks.

During preprocessing, we identified and removed 2 images with incorrect annotations, resulting in 998 images that were split randomly into 848 training images and 150 evaluation images. Additionally, the original images (840×1615 pixels) were cropped and resized to 320×640 pixels using bicubic interpolation to reduce GPU memory consumption and computational cost.

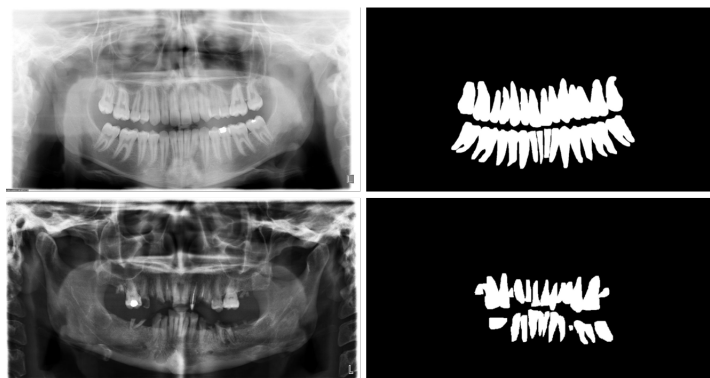


Figure 1. Illustration of image samples and their ground truth masks from the Tufts dental dataset.

3.1.2. UFBA-UESC Dental Dataset

The UFBA-UESC dental dataset [1] is comprised of 1500 labeled panoramic radiographs gathered at the University of Bahia. These images illustrate the variety of dental conditions commonly seen in patients at dental clinics and are classified in 10 categories defined in Table 1. Sample images and their associated segmentation masks are displayed in Figure 2.

Table 1. Categorization of the UFBA-UESC dental dataset images.

Category	Appliance	Restoration	Implant	All Teeth	>32 Teeth	Missing Teeth	Average Teeth	Number of Images
1	✓	✓		✓			32	73
2		✓		✓			32	220
3	✓			✓			32	45
4				✓			32	140
5			✓				18	120
6					✓		37	170
7	✓	✓				✓	27	115
8		✓				✓	29	457
9	✓					✓	28	45
10						✓	28	115

Note: ✓ indicates the presence of the corresponding characteristic in each category.

It is worth noting that the quality of some of the ground-truth annotations in Categories 4, 7, and 8 is inconsistent. In particular, several annotations exhibit coarse polygonal contours that do not accurately follow the anatomical boundaries of individual teeth (Figure 2). Nevertheless, these samples were retained because they are part of the original publicly available UFBA-UESC dataset used by previous studies. Despite these annotation imperfections, our experimental results demonstrate that the proposed model is capable of learning anatomically plausible tooth shapes. Consequently, all 1500 images were retained for training and evaluation. The original image dimensions (1991×1127 pixels) were

resized to 320×640 pixels using bicubic interpolation. The dataset was randomly divided into training and testing subsets. The same training and testing partition was used for all compared models to ensure a fair evaluation. No category-based stratification according to the categories presented in Table 1 was applied during the data partitioning. Approximately 80% of the images were used for training and 20% of images for testing.

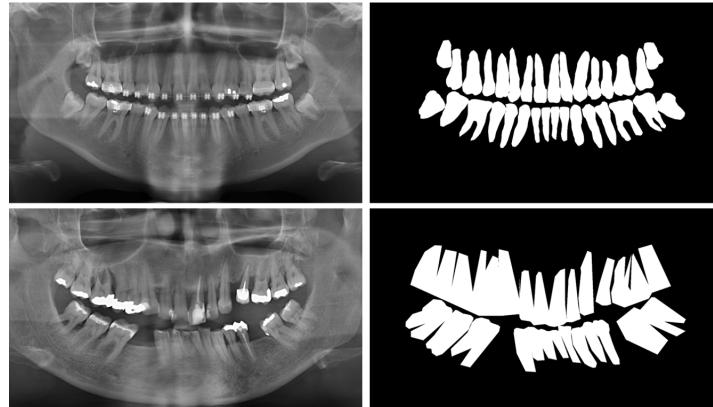


Figure 2. Illustration of image samples and their ground truth masks from the UFBA-UESC dental dataset.

3.2. The Proposed Architecture

DoubleU-Net [18] is an advanced variant of U-Net [19] that enhances segmentation performance by stacking two U-Net architectures enabling more detailed feature extraction. Its dual-stage architecture allows the second U-Net to refine predictions generated by the first, making it especially effective for complex segmentation tasks. Similarly, TransUNet [20], which combines the strengths of CNNs and Transformers, has demonstrated remarkable performance in medical image segmentation by leveraging global context and local feature extraction. Inspired by DoubleU-Net and TransUNet, we propose DoubleTransU-Net which incorporates comprehensive feature refinement and hybrid CNN-Transformer architecture to achieve higher precision and accurate segmentation.

3.2.1. TransUNet

TransUNet [20] is the first medical image segmentation framework that integrates the benefits of Transformers and U-Net [19], leveraging both fine-grained spatial features from CNNs and the global contextual information captured by Transformers. TransUNet employs a three-layer CNN to generate a feature map for the input. The Transformer then encodes image patch tokens obtained from the CNN feature map to capture global context. The self-attentive features encoded by Transformers are then upsampled, using a cascaded upsampler, to be combined with different high-resolution CNN features skipped from the encoding path, for enabling precise localization.

3.2.2. DoubleU-Net

DoubleU-Net [18] consists of two sequential U-Net [19] models, each comprising an encoder and a decoder. A pre-trained VGG-19 model [21], trained on ImageNet [22], is used in the first encoder, while the second encoder is constructed from scratch. After each encoder, Atrous Spatial Pyramid Pooling module (ASPP) [23] is employed to extract contextual features within each network. In the first network, the modified U-Net receives the input image, which produces a predicted mask. The input image and the produced mask undergo element-wise multiplication and the resulting output serves as an input for the second network that generates the final mask. The first decoder uses skip connections from the first encoder, while the second decoder incorporates skip connections from both

encoders, maintaining spatial resolution and enhancing output feature map quality. Additionally, a squeeze-and-excite block (SE) [24] is added in the first network's encoder and in the decoders of the first and second networks which enhances the feature maps quality by reducing the redundant information and passing the most relevant details.

3.2.3. Proposed Architecture: DoubleTransU-Net

DoubleTransU-Net uses two custom TransUNet [20] models in sequence. Similar to DoubleU-Net [18], the proposed architecture uses ASPP [23] modules and SE blocks [24]. An element-wise multiplication between the input image and the predicted mask produced by the first network is used to generate the input to the second network. Figure 3 presents the detailed architecture of DoubleTransU-Net.

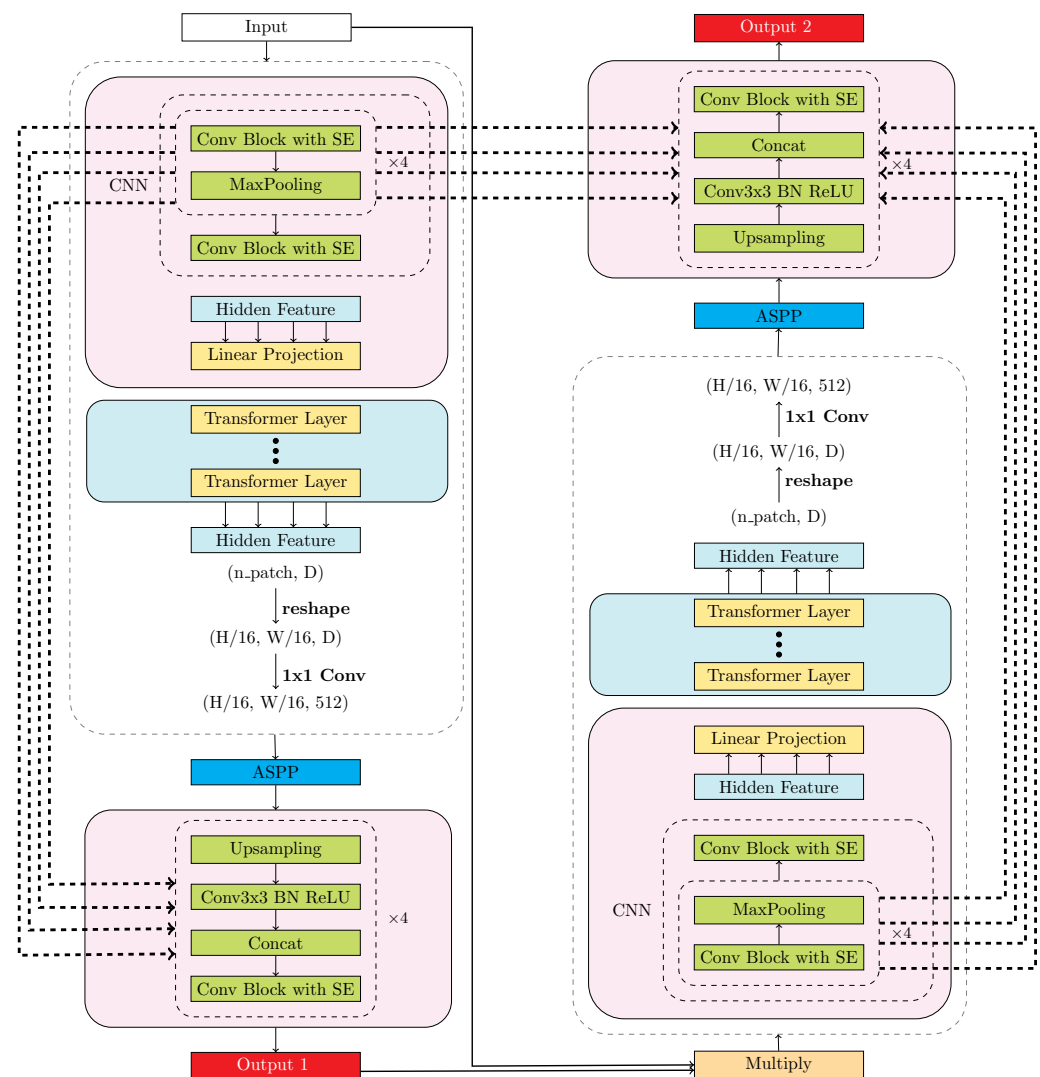


Figure 3. The proposed DoubleTransU-Net.

As illustrated in Figure 3, DoubleTransU-Net consists of two sequential TransUNet-based subnetworks. In the first stage, the input panoramic radiograph is processed by a CNN encoder enhanced with SE blocks, followed by Transformer layers that capture global contextual information. The resulting features are then passed through an ASPP module and a decoder to produce an initial tooth segmentation mask (Output 1). This predicted mask is multiplied element-wise with the original input image to emphasize the tooth regions while suppressing irrelevant background information before being forwarded to the second stage for further refinement. The second TransUNet follows the

same encoder–decoder structure and further refines the initial prediction. Unlike the first decoder, the second decoder receives skip connections from both encoders, enabling it to exploit complementary low-level spatial features extracted in both stages. The ASPP modules enrich the multi-scale contextual representation in each stage, while the SE blocks adaptively recalibrate channel responses to improve feature discrimination. The final decoder produces the refined segmentation mask (Output 2).

3.2.4. Encoder Architecture

In our proposed DoubleTransU-Net architecture, both networks utilize the same custom encoder from the TransUNet architecture. We modified TransUNet to handle inputs with dimensions where the width is twice the height, ensuring better preservation of the anatomical features specific to dental imagery.

The CNN is employed first to extract meaningful features from the input and produce a feature map. Each convolutional block includes two 3×3 convolutional layers, each followed by batch normalization to stabilize and accelerate training. After batch normalization, a Rectified Linear Unit (ReLU) activation function is applied to incorporate non-linear transformations. After that, an SE block is used to enhance the feature maps quality. After each block, a max-pooling operation is performed to reduce the spatial resolution by a factor of 2.

Next, the CNN feature map is reshaped into a sequence of flattened two-dimensional patches: $\{\mathbf{x}_p^i \in \mathbb{R}^{P^2 \cdot C} \mid i = 1, \dots, N\}$, where $P \times P$ denotes the spatial dimensions of each patch, and $N = \frac{HW}{P^2}$ denotes the total number of patches extracted from an input image.

Each vectorized patch $\mathbf{x}_p^i$ is projected into a D -dimensional latent embedding space through a learnable linear projection. To retain spatial information, positional embeddings specifically learned are added to the patch embeddings as follows:

$$\mathbf{z}_0 = [\mathbf{x}_p^1 \mathbf{E}, \mathbf{x}_p^2 \mathbf{E}, \dots, \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad (1)$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ represents the patch embedding projection, and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times D}$ is the position embedding.

The resulting tokens are subsequently processed through multiple Transformer layers, each comprising Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. The output of the ℓ -th layer can be formulated as:

$$\begin{aligned} \mathbf{z}'_{\ell} &= \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \\ \mathbf{z}_{\ell} &= \text{MLP}(\text{LN}(\mathbf{z}'_{\ell})) + \mathbf{z}'_{\ell}. \end{aligned} \quad (2)$$

where $\text{LN}(\cdot)$ represents the layer normalization operator and $\mathbf{z}_L$ corresponds to the encoded image representation.

After processing through multiple Transformer layers, the encoded feature size $\mathbf{z}_L \in \mathbb{R}^{\frac{HW}{P^2} \times D}$ is first reshaped from $\frac{HW}{P^2}$ to $\frac{H}{P} \times \frac{W}{P}$. Then we apply a 1×1 convolution to reduce the number of channels in the reshaped feature map to the number of classes. This reshaped 2D feature map contains enriched global and local information.

3.2.5. Decoder Architecture

The two decoders in DoubleTransU-Net utilize the same custom decoder architecture. Each block in the decoder performs an upsampling operation that doubles the spatial resolution of the features maps. Subsequently, a block of 3×3 convolution operation followed by batch normalization and ReLU activation function is applied. After skip connection concatenation, two 3×3 convolution operations are performed again, each followed by batch normalization and then by a ReLU activation function. Next, we employ

an SE block. Finally, a 1×1 convolutional layer with a sigmoid activation function is used to produce the final output mask. In the first decoder, we only use skip connections from the first encoder, but in the second decoder, we use skip connections from both the encoders, which preserve the spatial details and improve the quality of the output feature maps.

3.3. Loss Functions

3.3.1. IoU Loss Function

The IoU loss function is designed to maximize the overlap between the predicted segmentation mask and the ground truth mask. The IoU Loss is defined as:

$$\text{IoU Loss} = 1 - \frac{|P \cap G|}{|P \cup G|} = 1 - \text{IoU} \quad (3)$$

where P is the set of predicted pixels, and G is the set of ground truth pixels.

3.3.2. Dice Loss Function

Dice loss calculates the similarity between the predicted output and ground truth by focusing more on the overlap of positive pixels. The Dice loss is defined as:

$$\text{Dice Loss} = 1 - \frac{2 \times |P \cap G|}{|P| + |G|} = 1 - \text{Dice} \quad (4)$$

3.3.3. BCE Loss Function

The Binary Cross-Entropy (BCE) loss function is commonly used for binary classification problems. It quantifies the difference between the actual class labels (0 or 1) and the predicted probabilities output by the model. Mathematically, the formula for Binary Cross-Entropy loss is:

$$\text{BCE Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (5)$$

where N is the number of samples, y_i is the true label (either 0 or 1) and p_i is the predicted probability of the positive class (value between 0 and 1).

3.4. Implementation Details

Implementation was conducted using Python 3.11 with the TensorFlow framework on Google Colab Pro+, leveraging an NVIDIA A100 GPU with 40.0 GB of GPU RAM for accelerated computations. The environment provided 83.5 GB of system RAM and 235.7 GB of available disk storage, ensuring sufficient resources for the experiments. The training process utilized a batch size of 16, an initial learning rate of 0.001, and the Adamax optimizer. During the initial training stage, the training data were further divided into training and validation subsets. An EarlyStopping strategy was employed to monitor the validation performance and identify the approximate convergence point of each model. The learning curves consistently indicated convergence at approximately 30 epochs. Therefore, the validation subset was merged back into the training data, and all models were retrained on the complete training set for a fixed 30 epochs before being evaluated on the independent test set. Model performance was assessed using Accuracy, Dice coefficient, IoU, Recall, and Precision. The IoU loss function was used to optimize the models and improve segmentation performance.

4. Results

4.1. Comparison with State-of-the-Art Models

To evaluate the performance of our proposed DoubleTransU-Net model, we compared it against several deep learning-based methods such as U-Net [19], DoubleU-Net [18], Attention UNet [25], TransUNet [20] and DeepLabv3+ [26] on two publicly available datasets: Tufts [17] and UFBA-UESC [1] dental datasets. U-Net, DoubleU-Net and Attention UNet were implemented according to the original architecture described in the corresponding paper. DeepLabv3+ was trained with a Resnet50 [27] backbone and was pre-initialized with ImageNet-trained weights, while TransUNet utilizes the same CNN backbone and decoder as the one employed in our proposed architecture to allow a fair comparison.

4.1.1. Comparison on the Tufts Dental Dataset

Quantitative results on the Tufts dental dataset are presented in Table 2. DoubleTransU-Net achieves the highest overall segmentation performance among all compared methods, leading in Accuracy, Dice coefficient, IoU, and Recall while maintaining competitive Precision. Table 2 shows that DoubleTransU-Net's performance advantage over the single-stage baseline models is primarily reflected in Recall (93.97%) rather than Precision (92.27%, the third highest). This pattern suggests that the proposed model is more effective at recovering tooth regions that single-stage approaches fail to segment, thereby reducing false negatives while maintaining a competitive level of Precision. A comparison with DoubleU-Net further highlights this behavior. Although DoubleU-Net achieves the highest Precision (93.30%), its Recall (92.31%) is the lowest among the top-performing methods, indicating a more conservative segmentation behavior. In contrast, DoubleTransU-Net achieves a better balance between Precision and Recall, resulting in the highest Dice coefficient and IoU despite not attaining the highest Precision. These findings suggest that the combination of progressive refinement and Transformer-based global contextual learning enables DoubleTransU-Net to recover additional tooth regions without introducing a substantial increase in false-positive predictions, resulting in improved overall segmentation quality.

Table 2. Evaluation results (%) of different methods on the Tufts dental dataset.

Model	Accuracy	Dice	IoU	Precision	Recall
U-Net	91.72	92.48	86.01	92.02	93.33
TransUNet	91.76	92.62	86.25	92.51	93.09
DoubleU-Net	91.77	92.62	86.26	93.30	92.31
Attention UNet	91.52	91.79	84.82	92.38	91.56
DeepLabv3+	91.65	92.09	85.33	92.72	91.72
DoubleTransU-Net (Ours)	91.83	92.93	86.80	92.27	93.97

Note: Bold values indicate the highest result for each evaluation metric.

The qualitative comparisons shown in Figure 4 further support the quantitative results presented in Table 2. For the Tufts dataset (columns (a)–(c)), DoubleTransU-Net consistently demonstrates an improved ability to recover challenging fractured teeth that are either missed or only partially segmented by the competing methods. In column (a), DoubleTransU-Net is the only method that produces segmentation responses for both fractured teeth, although only small regions are recovered. In contrast, U-Net partially segments only the upper fractured tooth while introducing a false-positive fragment, TransUNet detects only the upper fractured tooth, and the remaining methods fail to identify one or both fractured structures. In columns (b) and (c), DoubleTransU-Net generally produces more complete segmentations of fractured teeth than the competing methods,

although one small fractured tooth in column (b) is better detected by DeepLabv3+, highlighting the difficulty of these particularly challenging cases. Notably, several tooth regions that remain undetected by TransUNet are at least partially recovered by DoubleTransU-Net. This observation is consistent with the quantitative improvements reported in the ablation study and suggests that enriching the first-stage TransUNet with ASPP and SE provides a more complete initial localization of challenging tooth structures. Since the first-stage prediction guides the refinement stage, detecting these regions early enables the second stage to further refine their segmentation rather than attempting to recover structures that were not localized initially.

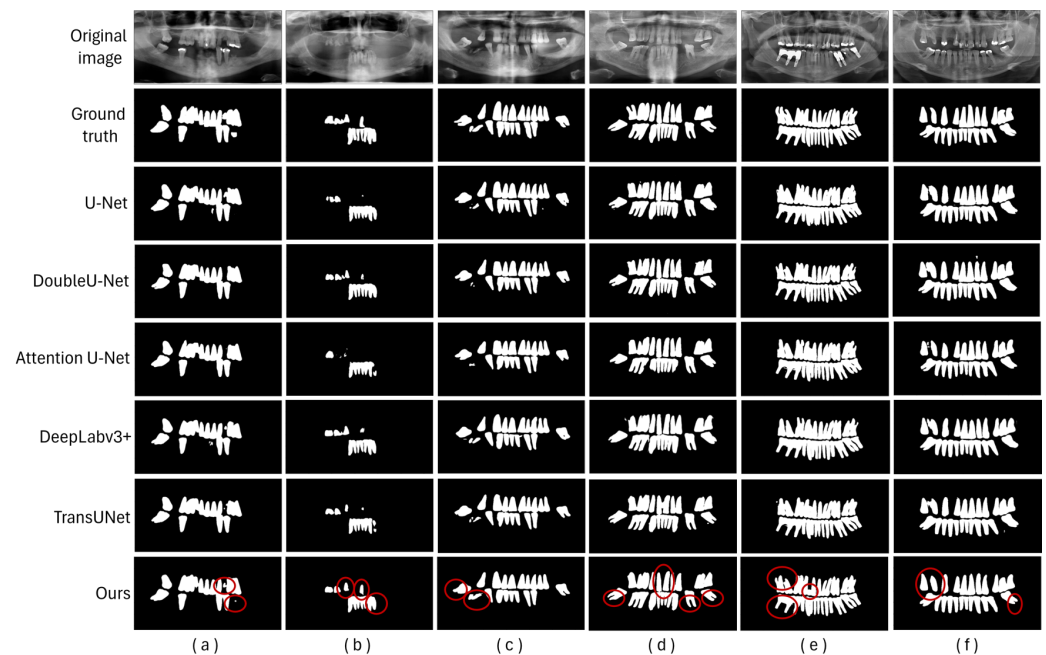


Figure 4. Qualitative visualization of teeth segmentation results. Columns (a–c) display the results obtained on the Tufts dental dataset, and columns (d–f) display the results obtained on the UFBA-UESC dental dataset.

4.1.2. Comparison on the UFBA-UESC Dental Dataset

Quantitative results on the UFBA-UESC dental dataset are presented in Table 3. DoubleTransU-Net achieves the highest overall segmentation performance among all compared methods, leading in Accuracy, Dice coefficient, and IoU while maintaining competitive Precision and Recall. On the UFBA-UESC dental dataset, the performance pattern differs from that observed on Tufts. Although DoubleU-Net achieves the highest Precision (94.73%) and U-Net the highest Recall (94.46%), DoubleTransU-Net attains the highest Dice coefficient and IoU without leading on either individual metric. This suggests that the proposed model achieves a more balanced trade-off between false-positive and false-negative predictions, resulting in improved overall segmentation quality. Given that the UFBA-UESC dataset comprises ten image categories with diverse dental conditions and imaging characteristics (Table 1), this balanced behavior may contribute to the model's robust performance across heterogeneous clinical scenarios.

Table 3. Evaluation results (%) of different methods on the UFBA-UESC dental dataset.

Model	Accuracy	Dice	IoU	Precision	Recall
U-Net	95.36	93.23	87.33	92.35	94.46
TransUNet	95.45	93.47	87.74	93.71	93.59
DoubleU-Net	95.50	93.50	87.79	94.73	92.62
Attention UNet	95.30	93.04	86.99	93.56	92.86
DeepLabv3+	95.38	93.20	87.26	92.98	93.62
DoubleTransU-Net (Ours)	95.54	93.72	88.18	93.57	94.24

Note: Bold values indicate the highest result for each evaluation metric.

The qualitative comparisons for the UFBA-UESC dental dataset are presented in columns (d)–(f) of Figure 4. These examples further support the quantitative results reported in Table 3 by demonstrating the proposed model’s ability to accurately delineate challenging tooth boundaries. Across columns (d)–(f), DoubleTransU-Net better preserves the anatomical morphology of multi-rooted teeth by maintaining clearer separation of the roots throughout the lower portion of the tooth. Owing to the intensity similarity between teeth, the alveolar bone, and surrounding tissues, competing methods tend to merge the roots into a single broad segmentation, producing only limited separation near the root tips. In contrast, DoubleTransU-Net better captures the natural root bifurcation, resulting in segmentation masks that more closely resemble the ground-truth annotations. In column (e), which contains multiple dental implants and a prosthetic bridge, DoubleTransU-Net accurately distinguishes the bridge from the adjacent natural teeth while simultaneously preserving the segmentation of the implants. Notably, DoubleU-Net also performs well in this case, suggesting that the progressive refinement strategy contributes to the accurate segmentation of complex prosthetic structures. Furthermore, column (f) contains a particularly challenging upper tooth exhibiting strong intensity similarity with the surrounding tissues. While most competing methods over-segment this structure, DoubleTransU-Net produces a segmentation mask that more closely matches the ground-truth annotation.

DoubleTransU-Net was also compared with recent work that used the same UFBA-UESC dental dataset. A summary of the results is provided in Table 4. Compared to models presented in other studies, our model’s performance in terms of Dice coefficient shows improvements of up to 1%, which indicates the capability of our method to enhance the accuracy of tooth segmentation in panoramic radiography.

Table 4. Performance comparison (%) of various methods on the UFBA-UESC dental dataset.

Model	Train Set	Test Set	Image Size	Loss Function	Accuracy	Dice	IoU
U-Net Ensembles [2]	1200	300	1127 × 1991	Tversky loss	95.21	93.63	-
TSASNet [3]	1200	150	-	Hybrid loss	96.94	92.72	-
MSLPNet [4]	1200	150	512 × 1024	Hybrid loss	97.30	93.01	-
DoubleTransU-Net (Ours)	1200	150	320 × 640	IoU loss	95.28	93.31	87.46
DoubleTransU-Net (Ours)	1200	300	320 × 640	IoU loss	95.54	93.72	88.18

4.2. Ablation Study

This section analyzes the impact of each individual component of our DoubleTransU-Net model and investigates the effect of different loss functions on overall performance on teeth segmentation by conducting an ablation study.

4.2.1. Ablation of Network Components

To understand the impact of our DoubleTransU-Net components, we performed a series of experiments on the Tufts dental dataset starting from TransUNet as a baseline model, that employs the same CNN backbone and decoder as employed in our proposed model to ensure a fair comparison. In the quantitative ablation in Table 5, SE denotes squeeze and excitation module, ASPP denotes Atrous Spatial Pyramid Pooling, and SkC denotes skip connection. TransUNet was first evaluated as-is, and then we added ASPP and SE modules to the model. After that we evaluated two baseline TransUNet architectures in sequence, then we evaluated adding skip connections from the first to the second network. After than we added ASPP module and finally, by adding SE module, our proposed model was evaluated in its entirety. Each added component improves segmentation performance, with the cumulative gain reaching 0.55% in IoU score and 0.31% in Dice coefficient over the baseline model.

Table 5. Ablation study on network components.

Model	Accuracy	Dice	IoU	Precision	Recall
TransUNet	91.76	92.62	86.25	92.51	93.09
TransUNet + ASPP + SE	91.77	92.71	86.41	92.10	93.71
2 × TransUNet	91.78	92.73	86.45	92.47	93.38
2 × TransUNet + SkC	91.81	92.79	86.55	93.17	92.79
2 × TransUNet + SkC + ASPP	91.82	92.87	86.69	91.97	94.16
DoubleTransUNet (Ours)	91.83	92.93	86.80	92.27	93.97

4.2.2. Ablation of Loss Functions

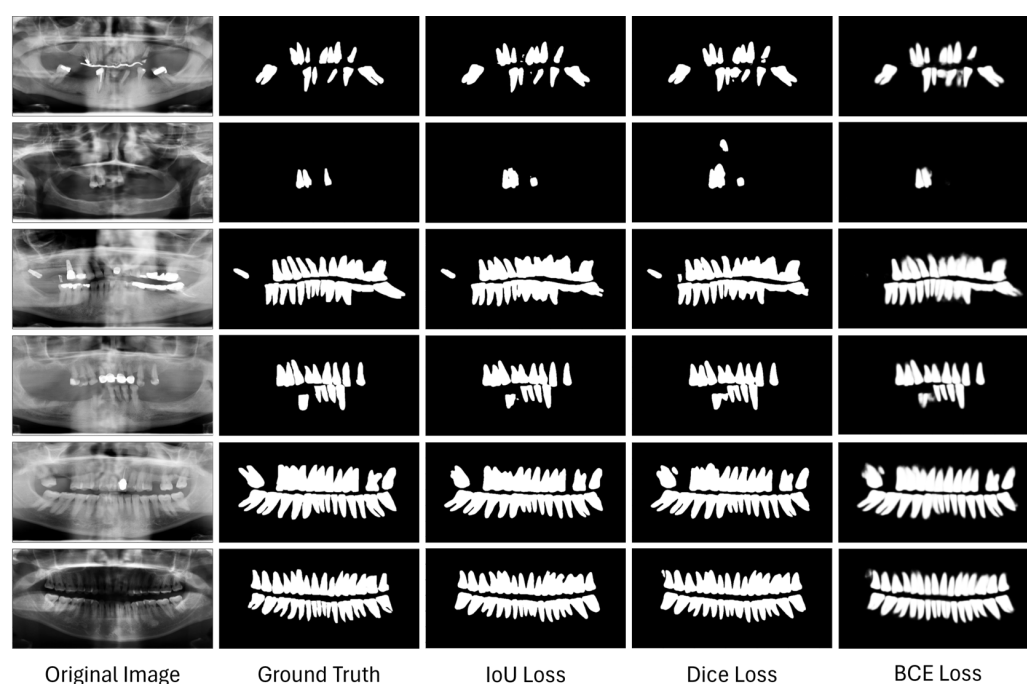
To further evaluate our approach and validate the choice of IoU loss as the loss function, experiments were carried out with multiple loss functions, including BCE loss and Dice loss. All loss functions were evaluated under identical conditions using our full proposed model. Table 6 displays the quantitative results, and a qualitative comparison is provided in Figure 5. The results indicate that the IoU loss function delivers significantly improved performance compared with BCE loss, with an improvement of 4.47% in IoU and 2.62% in Dice coefficient. While Dice loss also performs well, IoU loss consistently outperforms it in Accuracy, Dice coefficient, and IoU score. The qualitative comparisons presented in Figure 5 further explain these quantitative results. Models trained with BCE loss produce noticeably blurrier segmentation boundaries and exhibit both false-positive and false-negative predictions. For example, BCE loss fails to segment an existing implant in one case, misses an existing tooth in another, and incorrectly segments structures corresponding to a non-existent bridge and tooth. These observations are consistent with its substantially lower Dice coefficient and IoU. In contrast, the segmentation masks obtained using Dice loss are visually much closer to those generated with IoU loss, explaining the relatively small quantitative difference between the two loss functions. Nevertheless, several examples reveal that Dice loss occasionally predicts additional tooth regions that are absent from the ground-truth annotations, introducing false-positive segmentations despite otherwise accurate tooth delineation. These occasional errors explain why Dice loss consistently performs below IoU loss across all quantitative metrics. Overall, IoU loss produces segmentation masks that most closely resemble the ground-truth annotations, with more accurate boundary delineation and fewer false-positive predictions, providing visual evidence for its superior quantitative performance reported in Table 6.

Table 6. Ablation study of different loss functions.

Model	Accuracy	Dice	IoU	Precision	Recall
BCE Loss	91.79	90.31	82.33	91.78	91.96
Dice Loss	91.80	92.77	86.52	92.62	93.31
IoU Loss	91.83	92.93	86.80	92.27	93.97

4.3. Generalizability

To assess the generalizability of our proposed model, experiments were conducted using two public dental datasets: UFBA-UESC and Tufts dental datasets. All models were trained on the Tufts dental dataset (1000 images) and were evaluated on how well they generalize to unseen data from the UFBA-UESC dental dataset (1500 images). All models were evaluated using Accuracy, Dice coefficient, IoU, Recall and Precision. Table 7 summarizes the results of the generalizability experiments. DoubleTransU-Net achieved higher Dice and IoU scores compared to all state-of-the-art methods especially when compared to U-Net, DeepLabv3+ and Attention UNet.

**Figure 5.** Qualitative comparison of the different losses.**Table 7.** Generalizability experiment results.

Model	Accuracy	Dice	IoU	Precision	Recall
U-Net	95.07	92.19	85.56	91.05	93.76
TransUNet	95.16	92.47	86.04	90.96	94.43
DoubleU-Net	95.23	92.55	86.18	92.38	93.10
Attention UNet	95.07	92.15	85.48	91.59	93.11
DeepLabv3+	95.02	92.00	85.23	90.94	93.38
DoubleTransU-Net (Ours)	95.24	92.65	86.35	91.70	94.05

Note: Bold values indicate the highest result for each evaluation metric.

4.4. Model Complexity

The proposed DoubleTransU-Net architecture exhibits greater parametric complexity (67 M parameters) compared to baseline models such as U-Net (31 M) and TransUNet (23 M), owing to its dual-stage design incorporating sequential TransUNet networks augmented with ASPP modules and SE blocks. This increased computational burden is justified by three key advantages: (1) Superior segmentation performance, achieving state-of-the-art results of 86.80% in IoU score and 92.93% in Dice score on the Tufts dental dataset (representing improvements of 0.55% and 0.31% over TransUNet, respectively) and 88.18% in IoU score and 93.72% in Dice score on the UFBA-UESC dental dataset; (2) Enhanced capability in handling challenging cases including dental implants and fractured teeth, as demonstrated in qualitative results (Figure 4); and (3) Robust cross-dataset generalization, maintaining 86.35% in IoU and 92.65% in Dice performance when evaluated on unseen data (Table 7). While the model requires greater computational resources (~1949 s training time), this trade-off is clinically justified by its improved diagnostic utility through precise boundary delineation. Future architectural optimizations, such as network pruning or quantization, could potentially reduce computational overhead while preserving these performance advantages. The number of parameters and training times for different models on the Tufts dental dataset are reported in Table 8.

Table 8. Number of parameters and training times for different models on the Tufts dental dataset.

Model	Number of Parameters	Training Time (Seconds)
U-Net	31 M	530.23
TransUNet	23 M	766.56
DoubleU-Net	66 M	1501.06
Attention UNet	8 M	712.83
DeepLabv3+	17 M	408.48
DoubleTransU-Net (Ours)	67 M	1949.57

5. Discussion

Our proposed DoubleTransU-Net model outperformed all compared state-of-the-art methods in both IoU and Dice scores on two dental datasets. On the Tufts dental dataset, our model achieved improvements ranging from 0.54% to 1.98% in IoU score and from 0.31% to 1.14% in Dice score. On the UFBA-UESC dental dataset, the improvements range from 0.39% to 1.19% in IoU score and from 0.22% to 0.68% in Dice score. The results underscore the model's capacity to produce consistent segmentation results on panoramic dental X-rays. Furthermore, when compared with recent work that used the same UFBA-UESC dataset, our model outperformed other methods in Dice score despite using a substantially smaller input image size (320×640 versus 512×1024 or the original 1127×1991 resolution used in previous studies). Although downsampling inevitably reduces image resolution, these results suggest that the selected image size preserves sufficient anatomical information for semantic tooth segmentation while substantially reducing the computational cost. The qualitative results demonstrate that DoubleTransU-Net provides a more refined segmentation compared to other models. Teeth segmentation in panoramic radiography poses significant challenges, such as homogeneity in regions close to teeth, the presence of dental instruments and broken teeth. Despite these challenges, our model performs exceptionally well, producing segmentation masks that are closely aligned with the ground truth, as illustrated in Figure 4. Additionally, the generalizability experiments demonstrated the proposed model's robustness in handling unseen data. Our model achieved higher IoU

and Dice scores than all compared state-of-the-art models especially U-Net, DeepLabv3+ and Attention U-Net, with improvements of up to 1.12% and 0.65%, respectively.

Cropping the images of the Tufts dental dataset during preprocessing resulted in a gain in IoU and Dice scores, while the Accuracy dropped from 95.09% to 91.83%. However, in semantic segmentation, metrics like IoU and Dice score provide a stronger indication of the real performance of the model and the ablation study of loss functions results provide clear evidence. While BCE loss and IoU loss functions achieved similar Accuracy results with a difference of only 0.04%, the IoU and Dice scores results while using the IoU loss function are higher by 4.47% and 2.62%, respectively, compared to those achieved while using BCE loss function. The qualitative results shown in Figure 5 clearly highlight this difference.

In our model, we used a custom CNN backbone because it gave better performance across metrics compared to different pretrained backbones, and we utilized the same custom CNN backbone as well as the same decoder for the TranUNet model for a fair comparison. Since DoubleTransU-Net involves using two custom TransUNet architectures in sequence, it has a limitation in that it requires a large number of parameters, and while reducing the parameter count was considered, the ablation study shows that each component contributes significantly to the model's performance. As a result, retaining all components was necessary to achieve optimal results. On another hand, efforts were made to enhance the model's performance by experimenting with adding attention gates to skip connections, incorporating modules like CBAM [28] and applying various data augmentation techniques, but these modifications resulted in only marginal improvements. In future work, emphasis should be placed on enhancing the overall framework to improve teeth segmentation on dental X-rays by incorporating advanced preprocessing techniques, applying robust image enhancement methods and pretraining the model on dental X-rays.

6. Conclusions

In this work, we introduced the DoubleTransU-Net architecture to improve the semantic segmentation of teeth on dental panoramic radiographs. DoubleTransU-Net architecture uses its dual-stage design to refine features more effectively, making better use of the high-resolution spatial details captured by CNN features and the global context provided by Transformers. Additionally, it integrates ASPP module for capturing multi-scale contextual information and squeeze-and-excitation blocks for channel-wise attention, further enhancing its segmentation performance. The proposed model was assessed using two publicly accessible dental datasets and was compared against state-of-the-art models. The results demonstrated that our model consistently outperformed these baselines across various evaluation metrics. Furthermore, the model exhibited strong generalization capabilities, achieving robust performance on unseen data.

6.1. Limitations

Despite its strong segmentation performance, DoubleTransU-Net has three notable limitations. First, its dual-stage architecture substantially increases the model complexity compared with a single TransUNet, more than doubling the number of trainable parameters (67 M vs. 23 M) and increasing the average training time by approximately 2.5 times (1949.57 s vs. 766.56 s on the Tufts dataset, Table 8). Although this design contributes to improved segmentation accuracy, the large number of trainable parameters relative to the available training data also increases the potential risk of overfitting. While the proposed model demonstrated good generalization across two public datasets, larger training datasets would further improve confidence in the robustness of the learned representations. In addition, the increased model complexity results in higher computational and memory requirements, which may limit deployment in resource-constrained clini-

cal environments without further model compression, such as pruning, quantization, or knowledge distillation.

Second, the proposed framework performs semantic segmentation only, producing a binary tooth–background mask without distinguishing individual tooth instances or assigning tooth identifiers (e.g., FDI or Universal numbering systems). Consequently, it cannot be directly applied to clinical tasks that require instance-level segmentation or automatic tooth numbering, such as digital dental charting or tooth-specific pathology monitoring.

Third, the cross-dataset generalization study presented in Section 4.3 evaluates only one transfer direction, namely training on the Tufts dataset and testing on the unseen UFBA-UESC dataset. Although this provides evidence of the proposed model’s cross-dataset generalization capability, evaluation in a single transfer direction and on only two public datasets cannot fully characterize its robustness. A more comprehensive evaluation will require bidirectional cross-dataset experiments and validation on additional independent datasets acquired using different imaging protocols and patient populations.

6.2. Future Work

To further improve and extend this work, several research directions can be pursued. First, exploring instance-level segmentation could help enable more accurate separation of individual teeth, especially in cases with overlapping or closely positioned teeth. Additionally, the use of a larger and more precisely annotated dataset, together with improved regularization strategies, could further reduce the risk of overfitting while enhancing the model’s robustness and generalization across different patients or imaging conditions. Further optimization of the model architecture and training strategy may also lead to better performance. Lastly, the approach could be extended to other related tasks, such as the segmentation of dental pathologies (e.g., caries or periodontal lesions) expanding its utility in clinical and diagnostic applications.

Author Contributions: M.T. proposed the method, implemented the code, conducted the experiments, analyzed the results, and drafted the manuscript. A.B. supervised the work, provided critical feedback, and contributed to manuscript revisions. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The UFBA-UESC and Tufts panoramic dental radiograph datasets used in this study are publicly available from their respective sources. The original dataset publications report that the appropriate ethical approval was obtained prior to data collection. This study involved only the analysis of anonymized, publicly available data and therefore did not require additional ethical approval.

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

References

1. Silva, G.; Oliveira, L.; Python, M. Automatic Segmenting Teeth in X-ray Images: Trends, a Novel Data Set, Benchmarking and Future Perspectives. *Expert Syst. Appl.* **2018**, *107*, 15–31. [\[CrossRef\]](#)
2. 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.
3. 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. [\[CrossRef\]](#)
4. Chen, Q.; Zhao, Y.; Liu, Y.; Sun, Y.; Yang, C.; Li, P.; Zhang, L.; Gao, C. MSLPNet: Multi-Scale Location Perception Network for Dental Panoramic X-ray Image Segmentation. *Neural Comput. Appl.* **2021**, *33*, 10277–10291. [\[CrossRef\]](#)

5. Jader, G.; Fontineli, J.; Ruiz, M.; Abdalla, K.; Pithon, M.; Oliveira, L. Deep Instance Segmentation of Teeth in Panoramic X-ray Images. In *Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Paraná, Brazil, 29 October–1 November 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 400–407.
6. Silva, B.; Pinheiro, L.; Oliveira, L.; Pithon, M. A Study on Tooth Segmentation and Numbering Using End-to-End Deep Neural Networks. In *Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Virtual, 7–10 November 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 164–171.
7. Pinheiro, L.; Silva, B.; Sobrinho, B.; Lima, F.; Cury, P.; Oliveira, L. Numbering Permanent and Deciduous Teeth via Deep Instance Segmentation in Panoramic X-rays. In *Proceedings of the 17th International Symposium on Medical Information Processing and Analysis, Campinas, Brazil, 17–19 November 2021*; SPIE: Bellingham, WA, USA, 2021; Volume 12088, pp. 95–104.
8. Cui, W.; Zeng, L.; Chong, B.; Zhang, Q. ToothPix: Pixel-Level Tooth Segmentation in Panoramic X-ray Images Based on Generative Adversarial Networks. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Virtual, 13–16 April 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1346–1350.
9. Sheng, C.; Wang, L.; Huang, Z.; Wang, T.; Guo, Y.; Hou, W.; Xu, L.; Wang, J.; Yan, X. Transformer-Based Deep Learning Network for Tooth Segmentation on Panoramic Radiographs. *J. Syst. Sci. Complex.* **2023**, *36*, 257–272. [[CrossRef](#)] [[PubMed](#)]
10. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015*; pp. 3431–3440.
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; pp. 2961–2969.
12. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; pp. 8759–8768.
13. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid Task Cascade for Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019*; pp. 4974–4983.
14. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. ResNeSt: Split-Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022*; pp. 2736–2746.
15. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. PointRend: Image Segmentation as Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; pp. 9799–9808.
16. Rubiu, G.; Bologna, M.; Cellina, M.; Cè, M.; Sala, D.; Pagani, R.; Mattavelli, E.; Fazzini, D.; Ibba, S.; Papa, S.; et al. Teeth Segmentation in Panoramic Dental X-ray Using Mask Regional Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 7947. [[CrossRef](#)]
17. Panetta, K.; Rajendran, R.; Ramesh, A.; Rao, S.P.; Agaian, S. Tufts Dental Database: A Multimodal Panoramic X-ray Dataset for Benchmarking Diagnostic Systems. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 1650–1659. [[CrossRef](#)] [[PubMed](#)]
18. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In *Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Virtual, 28–30 July 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 558–564.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
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. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
22. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, USA, 20–25 June 2009*; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
23. Chen, L.C. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; pp. 7132–7141.
25. 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.
26. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; pp. 801–818.

27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

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.