A Comprehensive Study of Transformer-Based Segmentation for Dental Imaging using SwinUNET

Aiya Alchaar[1] AIYAJE@BU.EDU
Yuanbin Man[1] YBINMAN@BU.EDU
Saya Atchibay[1] SAYOKIT@BU.EDU
Farshid Alizadeh-Shabdiz[1] ALIZADEH@BU.EDU
[1] *Metropolitan College, Department of Computer Science, Boston University, MA, USA*
Peixi Liao[2] LIAOPX@BU.EDU
[2] *Henry M. Goldman School of Dental Medicine, Boston University, MA, USA*

Editors: Under Review for MIDL 2026

Abstract

Transformer-based architectures have reshaped the landscape of medical image segmentation, yet their application to high-resolution dental imaging remains underexplored. In this work, we study a Dual Swin-UNet (Lin et al., 2022) framework that leverages two complementary stages of Swin Transformer processing to more effectively segment dental structures. The method begins by dividing full-resolution dental radiographs into spatially consistent slices using a lightweight Swin-driven partitioning module. Each slice is then independently analyzed by a Swin-UNet decoder–encoder network, allowing the model to focus on localized anatomical details without sacrificing global context. Before segmentation, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance subtle boundaries, which yields nearly a 10% performance gain. After inference, all slice-level predictions are merged to reconstruct a complete segmentation map at the original resolution. Experimental results indicate that this two-stage design substantially improves segmentation quality compared to a standard Swin-UNet, particularly in challenging low-contrast regions. Our approach achieves state-of-the-art performance with a mean Dice score of up to 88% for tooth segmentation. These findings highlight the value of Transformer-based slicing strategies for detailed dental image analysis and demonstrate their potential to support clinical diagnostics and treatment planning.

Keywords: Swin-UNet, Dual Swin-UNet, Transformer-based Segmentation, Dental Image Segmentation, Medical Image Analysis, Multiclass Segmentation

1. Introduction

In this study, we focus on dental image segmentation to support clinical diagnosis. Radiographic (X-ray) images play a crucial role in dentistry by providing a comprehensive view of the oral and maxillofacial region, including the teeth, jaws, sinuses, and surrounding structures (Shah et al., 2014; Sergeeva et al.). However, accurately identifying individual teeth often requires detailed analysis and substantial clinical expertise. Automated diagnostic systems face challenges such as patient variability, image quality issues, differences in imaging devices, and noise (Ali, 2024). With advances in neural networks and deep learning, new methods now assist dentists in interpreting radiographs (Shan et al., 2021). Segmentation-based approaches, in particular, help partition panoramic images into

meaningful regions—such as teeth, bone, and sinuses—thereby simplifying interpretation. As deep learning continues to evolve, computer vision has become an essential component of modern medical image segmentation. The "U-Shaped" encoder-decoder architecture, particularly the U-Net model (Ronneberger et al., 2015), has achieved notable success in medical semantic segmentation. However, as image complexity and demands increase, U-Net alone may not meet all requirements. To address these challenges, researchers have developed various variant models, including UNet++ (Zhou et al., 2018), VNet (Milletari et al., 2016), ResUNet (Diakogiannis et al., 2020), TransUNet (Chen et al., 2021), and Swin-UNet (Cao et al., 2022). While these models have been developed using medium-sized open datasets, such as multi-organ and cardiac datasets, there is a scarcity of data specifically for dental images. Consequently, some models that perform well with such datasets may not be as effective for dental image segmentation tasks. Based on the state-of-the-art Swin-UNet model (Cao et al., 2022), which has been trained on the ImageNet (Deng et al., 2009) dataset with image sizes of 224 * 224 (SWin-Tiny) and 384 * 384 (SWin-Base) pixels, fitting dental images that may be 2k * 2k pixels requires resizing. However, resizing can lead to a loss of detail and potentially lower performance. To address this issue, a slice window approach was proposed before applying the Swin-UNet model. This approach preserves the spatial information of the images, thereby maintaining more detail, aiming to achieve state-of-the-art performance.

2. Related Work

CNN-Based Models: Early medical image segmentation relied on contour-based techniques and traditional machine learning methods. The introduction of deep CNNs, particularly U-Net (Ronneberger et al., 2015), transformed the field and inspired many extensions such as Res-UNet (Diakogiannis et al., 2020), Dense-UNet (Cai et al., 2020), U-Net++ (Zhou et al., 2018), and UNet3+ (Huang et al., 2020). U-Net concepts were also adapted to 3D tasks, leading to architectures like 3D-U-Net and V-Net (Milletari et al., 2016). Overall, CNN-based models continue to achieve strong performance in medical image segmentation due to their powerful feature representation capabilities. A recent study proposed by Xing et al. (Xing et al., 2024a,b) introduces an end-to-end pipeline that automates the segmentation of teeth and other oral structures from panoramic radiographs and facilitates implant planning. This method leverages ResUNet for the segmentation of dental structures, ensuring accurate delineation of teeth, bone, and other anatomical features. Following segmentation, the approach incorporates Principal Component Analysis (PCA) and linear regression to determine optimal implant locations. The Teeth U-Net model proposed by Hou et al. (Hou et al., 2023) is a deep learning segmentation model specifically designed for dental panoramic X-ray images. The model enhances traditional U-Net by incorporating a Multi-scale Aggregation attention Block (MAB) and a Dilated Hybrid self-Attentive Block (DHAB) at the bottleneck layer to address challenges in panoramic dental radiographs such as low contrast and intensity, variations in teeth shapes, and overlapping structures. Their results demonstrate that Teeth U-Net outperforms conventional U-Net and other segmentation models in identifying individual teeth and related anatomical structures with higher precision. While CNNs have been influential in advancing image segmentation, their reliance on local feature extraction limits their ability to capture global information. This

limitation has led to the exploration of transformer-based architectures.

Vision Transformers: The Transformer model was initially developed for machine translation tasks (Vaswani et al., 2017). Building on the success of Transformers, researchers introduced the Vision Transformer (ViT) (Dosovitskiy, 2020), which demonstrated a notable balance between speed and accuracy in image tasks. However, a significant limitation of ViT is its need for large datasets. To address this, the Data-efficient Image Transformer (Deit) framework proposed several training strategies that enable ViT to perform effectively on ImageNet (Touvron et al., 2021). Recently, advancements based on ViT have continued to emerge, including the introduction of the Swin Transformer. This efficient hierarchical vision Transformer, utilizing a shifted windows mechanism, has achieved state-of-the-art results in a range of vision tasks such as image classification, object detection, and semantic segmentation. Chen et al. (Sheng et al., 2023) proposed a Swin-UNet for tooth segmentation on panoramic radiographs, showing superior performance on the PLAGH-BH dataset compared with CNN-based models such as U-Net, LinkNet, and FPN. Ghafoor et al. (Ghafoor et al., 2023) introduced a Teeth Attention Block (TAB) integrated with an M-Net-like structure and Swin Transformer, achieving improved segmentation by better focusing on dental regions. These studies highlight the strength of ViT-based and attention-enhanced architectures for dental segmentation. Building on this, we employ the Swin Transformer block as the core of a U-shaped encoder–decoder with skip connections for segmentation.

3. Methodology

3.1. Dataset

Our study focuses on the 2D oral radiographs. The training and testing datasets consist of panoramic images, which generate a flat image of the curved jaw structure, typically highlighting the bones and teeth in detail. The initial phase of this research aims to develop a transformer model to segment the panoramic image into six categories: Bone, Major Nerve, Restoration, Root Canal, Sinus and Tooth. In addition, tangential radiograph images are also provided in our study, and they are used to capture specific views of the oral and maxillofacial region. Unlike conventional radiographs that offer a more general view, tangential radiographs focus on a particular area by projecting the X-ray beam tangentially to the structure of interest. By adding these images to the dataset, we aim to enhance data diversity and improve the model's decision-making capabilities.

The dataset includes annotations provided by dentists from Boston University Dental School and comprises a total of 200 images from 100 patients. Specifically, it consists of 100 panoramic and 100 tangential images, each type offering unique perspectives and details of the oral and maxillofacial region. These two types of images are utilized to fit the model.

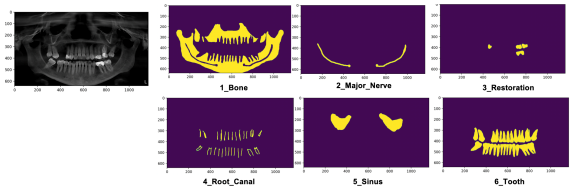

Figure 1: Panoramic Image with 6 categories.

3.2. Modeling

3.2.1. Dual Swin-Unet

Swin-UNET: Swin U-Net is a sophisticated model design that leverages the Swin Transformer blocks within a U-shaped encoder-decoder architecture for medical image segmentation. Here's a detailed description of the model design:
Encoder: 1) Input Layer: The input to the model is a medical image of a specific size (e.g., 224x224 or 384x384 pixels). 2) Swin Transformer Blocks: The encoder consists of multiple stages of Swin Transformer blocks. Each block is designed to handle different resolutions of the input image. Swin Transformer blocks are built using shifted windows, which partition the image into non-overlapping windows and compute self-attention within each window. This mechanism helps in capturing local context while also enabling efficient computation. 3) Hierarchical Feature Extraction: Each stage in the encoder progressively reduces the spatial dimensions of the feature maps (down sampling) while increasing the depth of the feature representations to capture multi-scale contextual information effectively.
Bottleneck: At the bottom of the U-shaped architecture, feature maps from the encoder are processed through additional Swin Transformer blocks to capture high-level features.
Decoder: 1) Upsampling Layers: The decoder consists of multiple upsampling layers that progressively increase the spatial dimensions of the feature maps. These upsampling layers are often implemented using transposed convolutions or interpolation methods.2) Skip Connections: It used to link corresponding layers of the encoder and decoder. These connections allow the model to utilize high-resolution features from the encoder directly in the decoding process, preserving spatial details. 3) Swin Transformer Blocks: Similar to the encoder, the decoder also incorporates Swin Transformer blocks to refine the upsampled feature maps and capture fine-grained details.
Output Layer: The final layer of the decoder produces the segmentation map, typically using a convolutional layer followed by a softmax or sigmoid activation function, depending on the nature of the segmentation task. In our study, the model outputs six classes.
Slicing Window: To preserve the original image information, particularly the spatial details, we designed a slicing windows approach. This technique addresses the issue of patch prediction, which often results in less smooth prediction masks compared to CNN results. While large-scale datasets typically improve performance, our study is constrained by a small domain dataset. To overcome this limitation, we split high-resolution images into non-overlapping 384 x 384 slices. These slices are then assembled into a large tensor and processed in parallel through the Swin UNet model. After processing, results are merged and restore to the original image size, ensuring that spatial information is retained and the final segmentation map is accurate. We present the entire pipeline in Figure 2.

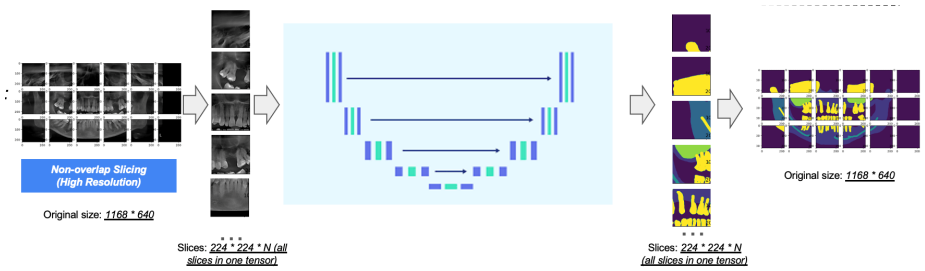

Figure 2: Dual Swin, Non-overlap slicing, fit to SWin-Unet, restore to original size.

3.2.2. Loss Function Selection

To fine-tune the Swin U-Net model, we modify the output layers by cutting them (including their weights) and appending new output layers tailored for our study's six-class segmentation task. For our study, we use the Swin Transformer with a base size of 384 × 384 as our backbone. This model, pretrained on the ImageNet 21K dataset, has over 71 million parameters. Despite the powerful pretrained backbone, it may not be well-suited to dental X-ray images. Therefore, we need to fine-tune all parameters using our specific dataset. When fine-tuning the model, we need to define loss functions for both the training and testing pipelines. We use the following loss functions: Dice Loss: This measures the overlap between the predicted and ground truth masks. The formula is:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{1}$$

where X and Y are the predicted and ground truth masks, respectively.
Categorical Cross-Entropy Loss: This measures the pixel-wise difference between predicted class probabilities and true labels, and is given by:

$$\mathcal{L}_{\text{CCE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i). \tag{2}$$

where C is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i. The total loss function is computed as an equal weighted sum of the components:

$$\mathcal{L}_{\text{total}} = 0.5\,\mathcal{L}_{\text{Dice}} + 0.5\,\mathcal{L}_{\text{CCE}}. \tag{3}$$

Here, Dice Loss and CCE Loss are weighted equally, with each contributing 50% to the total loss. In addition, we use the mean Intersection over Union (mIoU) to evaluate the model's performance. For the learning rate, we start with an initial value of 0.005 and use a power decay schedule. We use Stochastic Gradient Descent (SGD) as the optimizer.

3.2.3. Hyperparameter Tuning

Hyperparameter tuning plays a crucial role in optimizing deep learning models, particularly in medical imaging, where input images are often high-resolution, leading to significant memory constraints. Among the various hyperparameters, learning rate and batch size have a considerable impact on model convergence and performance (Bengio, 2012). In this study, we conducted an extensive search for optimal hyperparameters, evaluating learning rates in the range of 10e-3 to 10e-7 using a power decay schedule in all cases, and batch sizes of 4, 8, and 16. Due to the computational constraints associated with medical image segmentation, lower batch sizes are commonly used to balance memory consumption and performance. We evaluated model performance using Dice similarity coefficient, Intersection over Union, and precision. Our results, visualized in Figure 3, demonstrate a clear trend: models trained with smaller batch sizes and larger learning rates tend to exhibit better performance. Higher learning rates facilitate faster convergence by enabling larger weight updates, whereas smaller batch sizes contribute to improved generalization by introducing more variance in gradient updates. Conversely, models trained with larger batch sizes and

lower learning rates often converge more slowly and may struggle to escape sharp local minima, leading to suboptimal performance. These findings highlight the importance of tuning hyperparameters to achieve optimal segmentation accuracy.

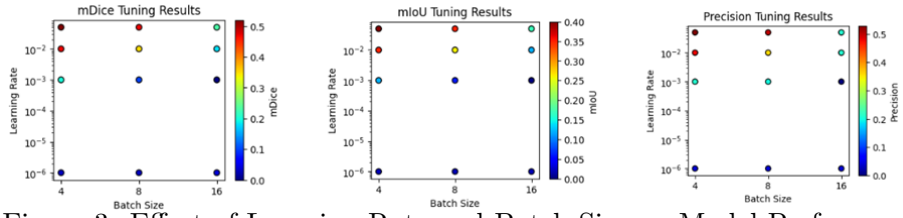

Figure 3: Effect of Learning Rate and Batch Size on Model Performance.

4. Experiments

4.1. Implementation Details

Our experiment is based on Python 3.10 and PyTorch 2.2. The dataset consists of 200 panoramic and tangential x-ray images, split into 80% for training and 20% for testing. To evaluate the impact of data augmentation, we conducted experiments under two settings: one with no augmentation, using the original dataset, and another with augmentation, applying horizontal flipping and slight 5-degree rotations. The input image size is set to 384 $\times$ 384 pixels, with a patch size of 4. We train our model on Google Colab using an Nvidia A100 GPU with 32GB of memory. We initialize the model parameters with weights pre-trained on ImageNet but replace the output layers with new ones suited to our task. During training, we employ the SGD optimizer with a momentum of 0.9 and weight decay of 1e-4 for optimization, for 150 epochs.

4.2. Baseline Experiment Results

The comparison of the proposed Dual Swin U-Net with previous models on the dental dataset is presented in Figure 4. The testing results indicate that the slicing window approach outperforms the original model, leading to improved accuracy and smoother segmentation boundaries. This improvement can be attributed to the localized feature extraction enabled by the slicing window technique, which helps the model capture intricate structures.

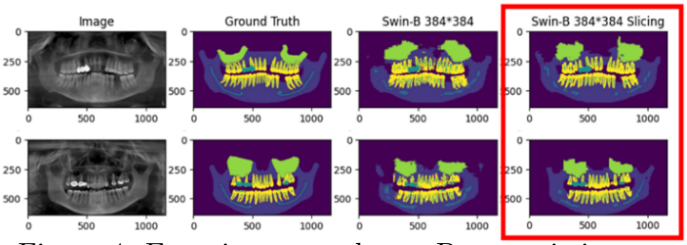

Figure 4: Experiment results on Panoramic images.

To enhance performance, we incorporate data augmentation and compare model predictions on augmented and non-augmented datasets. Figure 5 presents a sample prediction using the augmented dataset, demonstrating clearer segmentation boundaries and improved

class distinction. The results indicate that the model benefits from increased data diversity while maintaining anatomical consistency.

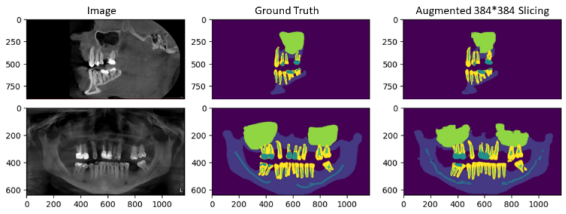

Figure 5: Slicing sample prediction using the augmented dataset

For a comprehensive evaluation, we report both overall and class-wise performance. Tables 1–3 summarize the results across different experimental settings. Table 1 shows performance on the original dataset, while Table 2 highlights the improvements gained through data augmentation. Table 3 reports per-class performance, highlighting variations in segmentation accuracy among the six classes. Overall, larger input resolutions improve segmentation quality, with 384×384 outperforming 192×192. Among the classes, "teeth" and "bone" achieve the highest accuracy. Non-overlap slicing also performs better than overlap slicing, indicating that preserving independent spatial regions benefits segmentation. Data augmentation further boosts performance, yielding more accurate boundaries and higher overall metrics, demonstrating that increased data diversity enhances generalization.

Table 1: Overall Performance

Dataset	**mDice**	**hd95**	**mIoU**	**Prec**
(192, 192, Swin-B)	0.47	79.1	0.35	0.50
(Slice192, Slice192, Swin-B, Non-Overlap)	0.58	67.4	0.45	0.60
(Slice192, Slice192, Swin-B, Overlap (96, 96))	0.41	232	0.30	0.44
(384, 384, Swin-B)	0.54	72.4	0.41	0.54
(Slice384, Slice384, Swin-B, Non-Overlap)	0.60	68.5	0.47	0.61
(Slice384, Slice384, Swin-B, Overlap (192, 192))	0.53	212	0.41	0.52

Table 2: Augmented Dataset Performance

Dataset	**mDice**	**hd95**	**mIoU**	**Prec**
Overall Performance				
(200, 384, 384, Swin-B)	0.57	70.1	0.45	0.59
(200, Slice384, Slice384, Swin-B, Non-Overlap)	0.62	62.9	0.50	0.65
Class-Wise Performance				
(Slice384, Slice384, Non-Overlap) – Bone	0.81	66.4	0.69	0.79
(Slice384, Slice384, Non-Overlap) – Major Nerve	0.32	61.9	0.21	0.40
(Slice384, Slice384, Non-Overlap) – Restoration	0.64	69.5	0.53	0.70
(Slice384, Slice384, Non-Overlap) – Root Canal	0.51	51.0	0.35	0.53
(Slice384, Slice384, Non-Overlap) – Sinus	0.65	92.4	0.54	0.69
(Slice384, Slice384, Non-Overlap) – Tooth	0.79	36.4	0.66	0.78

Table 3: Class-Wise Performance

Dataset	Class	mDice	hd95	mIoU	Prec
(192, 192, Swin)	Bone	0.71	82.9	0.56	0.70
(192, 192, Swin)	Maj. Nerve	0.17	85.8	0.12	0.23
(192, 192, Swin)	Restoration	0.40	129	0.28	0.46
(192, 192, Swin)	Root Canal	0.28	70.8	0.16	0.31
(192, 192, Swin)	Sinus	0.64	68.7	0.51	0.68
(192, 192, Swin)	Tooth	0.65	37.3	0.49	0.62
(Slice192, Slice192, Non-Overlap)	Bone	0.78	90.9	0.64	0.75
(Slice192, Slice192, Non-Overlap)	Maj. Nerve	0.24	68.3	0.15	0.29
(Slice192, Slice192, Non-Overlap)	Restoration	0.61	76.1	0.48	0.63
(Slice192, Slice192, Non-Overlap)	Root Canal	0.45	45.6	0.29	0.48
(Slice192, Slice192, Non-Overlap)	Sinus	0.65	97.5	0.52	0.68
(Slice192, Slice192, Non-Overlap)	Tooth	0.76	26.2	0.62	0.75
(Slice192, Slice192, Overlap (96, 96))	Bone	0.68	200	0.52	0.69
(Slice192, Slice192, Overlap (96, 96))	Maj. Nerve	0.06	504	0.03	0.05
(Slice192, Slice192, Overlap (96, 96))	Restoration	0.29	302	0.18	0.21
(Slice192, Slice192, Overlap (96, 96))	Root Canal	0.31	91.2	0.19	0.25
(Slice192, Slice192, Overlap (96, 96))	Sinus	0.41	273	0.30	0.60
(Slice192, Slice192, Overlap (96, 96))	Tooth	0.72	24.1	0.57	0.81
(384, 384, Swin)	Bone	0.77	75.3	0.63	0.73
(384, 384, Swin)	Maj. Nerve	0.25	81.1	0.16	0.29
(384, 384, Swin)	Restoration	0.49	113	0.37	0.50
(384, 384, Swin)	Root Canal	0.35	59.7	0.22	0.36
(384, 384, Swin)	Sinus	0.66	70.2	0.55	0.66
(384, 384, Swin)	Tooth	0.71	34.9	0.55	0.70
(Slice384, Slice384, Non-Overlap)	Bone	0.80	79.9	0.68	0.77
(Slice384, Slice384, Non-Overlap)	Maj. Nerve	0.27	83.7	0.17	0.33
(Slice384, Slice384, Non-Overlap)	Restoration	0.61	60.5	0.49	0.67
(Slice384, Slice384, Non-Overlap)	Root Canal	0.48	54.6	0.32	0.50
(Slice384, Slice384, Non-Overlap)	Sinus	0.65	103	0.53	0.66
(Slice384, Slice384, Non-Overlap)	Tooth	0.77	30.0	0.62	0.76
(Slice384, Slice384, Overlap (192, 192))	Bone	0.79	117	0.66	0.79
(Slice384, Slice384, Overlap (192, 192))	Maj. Nerve	0.14	505	0.10	0.13
(Slice384, Slice384, Overlap (192, 192))	Restoration	0.46	271	0.34	0.39
(Slice384, Slice384, Overlap (192, 192))	Root Canal	0.38	88.5	0.24	0.36
(Slice384, Slice384, Overlap (192, 192))	Sinus	0.62	263	0.50	0.67
(Slice384, Slice384, Overlap (192, 192))	Tooth	0.78	26.0	0.65	0.79

4.3. Refinement Experiments with CLAHE Preprocessing

After performing baseline experiments and identifying that the best performance was achieved using cropped slices of size 384×384 with no overlap, we conducted further experiments based on this setting to enhance model performance. We applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to all images prior to training. CLAHE enhances the local contrast of images by applying histogram equalization to small image regions (tiles), with a limit on contrast amplification to avoid amplifying noise. This is especially useful in medical images where soft tissue structures, such as nerves or sinuses, often have low contrast against the background. In our pipeline, CLAHE was applied to the full-resolution

images prior to the slicing step. Sample enhanced images are presented below to illustrate the effect of this preprocessing technique. The top row shows images after applying CLAHE preprocessing, while the bottom row shows the corresponding original images.

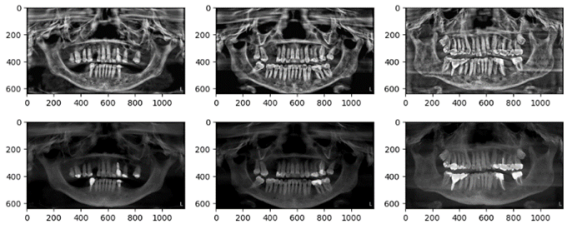

Figure 6: CLAHE (top row) compared to the original images (bottom row).

Additionally, we restricted the training to a subset of classes: bone, nerve, teeth, and sinus. These four classes were selected due to their consistent presence and clinical relevance in dental diagnostics. To simplify the class structure while retaining valuable information, we merged the restoration class into the teeth class, as restorations are commonly treated as part of the tooth in clinical settings. Less frequent classes, such as root canal, were excluded to reduce label noise and enhance overall model stability. The performance of the model under this refined setup improved notably in several classes. Below are examples of predicted segmentations from the refined four-class configuration with CLAHE, followed by a table summarizing the quantitative results.

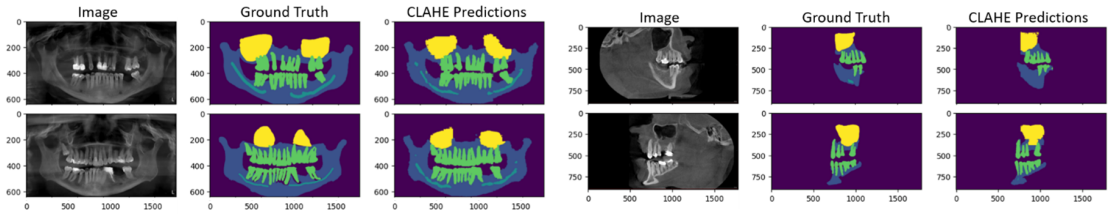

Figure 7: Example segmentation results for the refined four-class configuration.

Table 4: Enhanced Labels with CLAHE Results

Dataset	**mDice**	**hd95**	**mIoU**	**Prec**
Overall Performance				
(200, Slice384, Slice384, Swin-B, Non-Overlap)	0.68	61.5	0.58	0.70
(200, Slice384, Non-Overlap) Maj. Nerve Excluded	0.81	64	0.71	0.81
Class-Wise Performance				
(Slice384, Slice384, Non-Overlap) – Bone	0.83	51	0.72	0.81
(Slice384, Slice384, Non-Overlap) – Major Nerve	0.29	54	0.20	0.35
(Slice384, Slice384, Non-Overlap) – Tooth + Restor.	0.88	40	0.79	0.87
(Slice384, Slice384, Non-Overlap) – Sinus	0.72	101	0.61	0.76

Compared to the original multi-class setup, the refined configuration led to notable improvements in segmentation performance. The teeth class (merged with restorations) saw a substantial increase in mDice (from 0.77 to 0.88) and mIoU (from 0.62 to 0.79), indicating

that merging restorations into the teeth class simplified the learning task and enhanced boundary detection. The bone class also showed consistent improvement across all metrics, particularly a sharp reduction in Hausdorff Distance (from 79.9 to 51). Moreover, the bone mIoU increased after enhancement (from 0.68 to 0.72). For sinus, contrast enhancement and label correction led to a measurable gain in mDice (from 0.65 to 0.72) and improved precision. Overall, the results show that focusing on clinically significant classes combined with careful preprocessing can yield more reliable performance.

5. Analysis and Discussion of Results

The experimental results provide valuable insights into the impact of image resolution, slicing strategies, and overlap conditions on segmentation performance using Swin-UNet.
1) Comparison of Patch Sizes: 192 vs. 384: A comparison between the two different patch sizes, 192×192 and 384×384, reveals that the larger patch size consistently achieves superior segmentation performance. Specifically, the mDice and mIoU scores for 384 × 384 patches are higher than those of 192×192 across experiments. This trend suggests that using larger patches allows the model to capture more spatial context, which is particularly beneficial for segmenting complex anatomical structures such as teeth.
2) Effect of Slicing on Segmentation Performance: Results show that applying slicing techniques significantly enhances segmentation performance across all evaluation metrics. Both mDice and mIoU scores improved when slicing was introduced, compared to training on the original resized images. This improvement can be attributed to the model focusing on finer details within each patch, leading to better feature extraction and segmentation accuracy. The increase in precision also suggests that slicing reduces misclassification by allowing the model to make more confident predictions.
3) Impact of Overlap vs. Non-Overlap Slicing: Among the different slicing strategies, non-overlapping slicing outperforms overlapping slicing in almost all cases. For instance, with slice size 384×384, the non-overlapping approach achieved higher mDice and mIoU scores (0.80/0.68 for bone and 0.61/0.49 for restorations) compared to the overlapping setting (0.79/0.66 and 0.46/0.34, respectively). Similarly, for the 192×192 configuration, non-overlapping slices yielded superior results (mDice/mIoU of 0.78/0.64 for bone and 0.76/0.62 for tooth) relative to overlapping slices (0.68/0.52 and 0.72/0.57, respectively). These results indicate that overlapping patches introduce redundancy and inconsistencies during merging, while also increasing computational overhead without improving performance.

6. Conclusion

In this paper, we investigated a transformer-based U-shaped architecture with a slicing windows approach for dental segmentation. By integrating the Swin U-Net with the slicing methodology, our model achieves state-of-the-art performance, albeit with higher GPU requirements compared to standard Swin models. Our results highlight the importance of slicing as a preprocessing step. Larger patches (384×384) capture more contextual information and outperform smaller ones, while non-overlapping slicing consistently delivers the best performance across metrics and reduces computational overhead. These findings offer valuable guidance for optimizing deep learning-based dental image segmentation.

Acknowledgments

We would like to express our sincere gratitude to the dental department for their valuable insights and for providing access to the dataset, which was essential for this research. Additionally, we extend our appreciation to our advisors from the computer science department for their guidance and constructive feedback, which greatly contributed to refining our methodology and analysis. Their support and expertise have been instrumental in the successful completion of this work.

References

Magdi A Ali. The role of artificial intelligence in modern dentistry: Applications, challenges, and future directions. *Future Dental Research*, 2(2):39–49, 2024.

Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade: Second edition*, pages 437–478. Springer, 2012.

Sijing Cai, Yunxian Tian, Harvey Lui, Haishan Zeng, Yi Wu, and Guannan Chen. Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, 10(6):1275, 2020.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Afnan Ghafoor, Seong-Yong Moon, and Bumshik Lee. Multiclass segmentation using teeth attention modules for dental x-ray images. *IEEE Access*, 11:123891–123903, 2023.

Senbao Hou, Tao Zhou, Yuncan Liu, Pei Dang, Huiling Lu, and Hongbin Shi. Teeth u-net: A segmentation model of dental panoramic x-ray images for context semantics and contrast enhancement. *Computers in Biology and Medicine*, 152:106296, 2023.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. Ieee, 2020.

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

II Sergeeva, TF Tikhomirova, VV Rozhkovskaya, and NA Savrasov. Methods of x-ray examination and x-ray diagnostics in dentistry. x-ray methods of research.

Naseem Shah, Nikhil Bansal, and Ajay Logani. Recent advances in imaging technologies in dentistry. *World journal of radiology*, 6(10):794, 2014.

T Shan, FR Tay, and L Gu. Application of artificial intelligence in dentistry. *Journal of dental research*, 100(3):232–244, 2021.

Chen Sheng, Lin Wang, Zhenhuan Huang, Tian Wang, Yalin Guo, Wenjie Hou, Laiqing Xu, Jiazhu Wang, and Xue Yan. Transformer-based deep learning network for tooth segmentation on panoramic radiographs. *Journal of systems science and complexity*, 36 (1):257–272, 2023.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yang Xing, Peixi Liao, Reem AwdhE Alasleh, Vissuta Khampatee, and Farshid Alizadeh-Shabdiz. Dental x-ray segmentation and auto implant design based on convolutional neural network. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 243–246. IEEE, 2024a.

Yang Xing, Peixi Liao, Reem AwdhE Alasleh, Vissuta Khampatee, and Farshid Alizadeh-shabdiz. X-ray segmentation and implant design using panoramic and tangential views. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, pages 260–265, 2024b.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.