Applications of Foundation Models and Efficient Diffusion Models to Dental Radiography to Assist Dental Treatment Planning

Matthew Lewis
University of Central Florida
Orlando, FL
matthew.lewis@ucf.edu

Melody Halbert
University of Central Florida
Orlando, FL
melody.halbert@ucf.edu

Laura Brattain
University of Central Florida
Orlando, FL
laura.brattain@ucf.edu

Abstract

Dental radiographs are an essential diagnostic tool for identifying oral pathologies, missing teeth, and maxillofacial structures. Manual interpretation of these radiographs can be time-consuming, inconsistent, and prone to human error, motivating the development of automated approaches for tooth detection, classification, and reconstruction. In this study, we present a multi-stage pipeline combining object detection, segmentation, classification, and generative modeling to automate tooth identification and synthesis in dental radiographs. We evaluate three segmentation methods: fine-tuned MobileSAM, YOLOv8 segmentation, and YOLOv8 bounding boxes refined with MobileSAM. These are evaluated on two dental datasets and demonstrate that YOLOv8 achieves superior segmentation performance across Dice, IoU, and recall metrics. We further implement post-processing heuristics to enforce anatomical plausibility, improving classification accuracy from 0.951 to 0.967 by correcting mirrored and duplicate tooth predictions. Finally, we train a diffusion model on cropped tooth images to generate anatomically realistic reconstructions of missing teeth, demonstrating the feasibility of high-quality synthetic tooth generation. These results highlight the potential of integrating detection, classification, and diffusion-based synthesis to advance automated dental radiograph analysis and lay the groundwork for future clinically oriented dental AI systems.



Figure 1. Example dental radiograph.

1. Introduction

Dental radiographs, or dental x-rays, are used to analyze features and deformities that wouldn't otherwise be noticeable in route oral exams and inspection (Fig. 1). These images can pinpoint issues such as cavities, restorations, implants, and impacted teeth, while also highlighting other maxillofacial anatomical structures. Accurate and careful radiographic interpretation is essential for the diagnosis and treatment of dental diseases. In particular, as the result of severe cavities, gum diseases, or accidents (e.g. sports or trauma), missing teeth are common in dental x-rays; however, manual charting of missing tooth spaces can be inconsistent between dental professionals (dentists, dental hygienists, dental assistants, etc.). Manual workflow can also be time intensive, leading to inaccurate documentation, especially in high-volume clinics or time-sensitive settings.

There is a need to automate the charting of both the present and missing teeth for clinical applications. Automating this step would improve communication, improve patient safety, and optimize clinical workflows. Models for accurately identifying tooth numbers would be a key component in multi-step dental AI systems. Selective realistic modification of dental radiographs such as generation of synthetic teeth to fill in the empty spaces, adding pathology or incorporating potential treatment options (implants, restorations, root canals, etc.) would be a valuable tool to generate robust digital libraries for future development of dental AI tools and to perform AI Interpretability research. In addition, this technology could aid patient education, treatment planning, and dental education.

2. Related Works

2.1. Real-time Object Detection

Faster R-CNN is a convolutional neural network for object detection. It uses a Region Proposal Network (RPN) to create a set of rectangular object proposals, which are then fed into Fast R-CNN, the previously state-of-the-art object detection network [1, 10]. Faster R-CNN proves to have state-of-the-art accuracy on MS COCO and PASCAL VOC, two commonly used datasets for object detection.

Mask R-CNN furthers the goal of Faster R-CNN by adding a third output of an object mask [3]. This is done by predicting segmentation masks on each Region of Interest in parallel with the preexisting Mask R-CNN branches. Mask R-CNN adds the ability to detect bounding boxes and objects by each pixel.

You Only Look Once (YOLO) improved on this by approaching the image at once split into smaller grids. YOLOv8 improves on this idea with better performance, flexibility, and efficiency [12]. YOLOv8 was chosen due to its extensive community adoption and support, its strong segmentation performance in medical tasks and domains beyond natural image datasets, and its ability to be fine-tuned effectively on relatively small datasets.

2.2. Diffusion Models

Variational Auto-Encoders (VAEs) are probabilistic generative models that use a neural network to encode input data into a latent distribution with reduced memory cost [5]. With VAEs, the model can understand the latent distribution itself and refer to specific features as coordinates along a continuous vector.

Generative Adversarial Networks (GANs) improve on the images generated by VAEs and similar papers by introducing Adversarial Training [2]. In Adversarial Training, the generator creates synthetic samples which are then passed to a discriminator with real data which is tasked with distinguishing the real data from fake. This process trains the network to create increasingly realistic data.

Denoising Diffusion Probabilistic Models (DDPMs) reframe generative modeling as an iterative denoising process [4]. By progressively adding noise and denoising to train, DDPMs achieve high quality and diverse results without the need for adversarial training.

Instead of running directly on the pixel space like DDPMs, Stable Diffusion converts images into a latent space using VAEs for the diffusion process [11]. This method not only improves the quality and diversity of results from previous diffusion methods, but also allows for a variety of extensions to be used in tandem including conditioning and inpainting. Stable Diffusion was chosen due to its fast inference time and low memory cost.

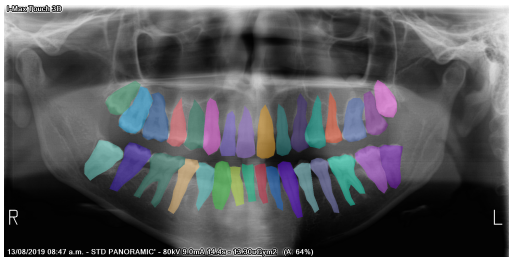


Figure 2. Example of a dental radiograph with its ground truth segmentation overlaid.



Figure 3. Example of a pediatric radiograph excluded from the dataset due to the presence of baby teeth beyond the standard 1–32 Universal Numbering System.

3. Materials & Methods

3.1. Dataset

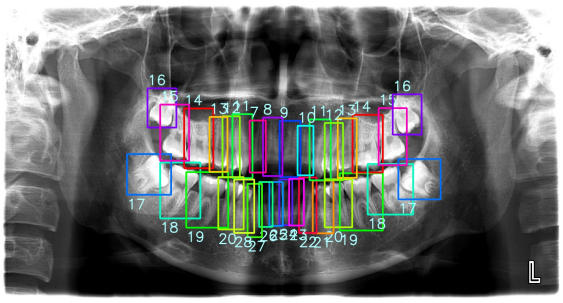
We used two datasets for this study, both dental radiographs with ground truth segmentations and labellings from 1-32 as is standard in the Universal Numbering System. The first dataset contains 598 images taken in Asuncion, Paraguay [7, 8]. The second dataset contains 1000 images from various hospitals [9]. Both datasets were segmented and labeled by a team of researchers (Fig. 2). For our needs, in both datasets, we removed any pediatric images (where baby teeth other than 1-32 are shown; see Fig. 3). Using both datasets allowed for increased variability and improved generalization of the model.

3.2. Computing Resources

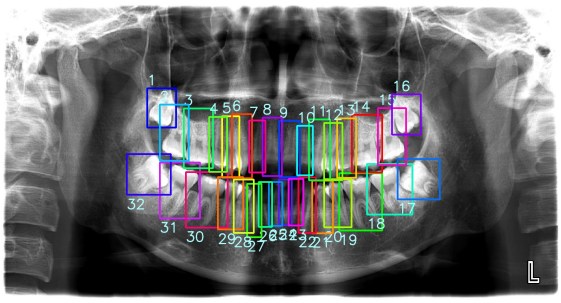
To train our model and perform inference, we utilized the Newton GPU Cluster at the UCF Advanced Research Computing Center (ARCC) Our model was trained utilizing NVIDIA Tesla V100 or H100 GPUs.

3.3. Segmentation

Finding an accurate, fast, and computationally efficient segmentation method was crucial for this study. We wanted something that could easily be implemented in dental centers across the globe, including those in resource-limited settings. After deliberation, we decided on testing MobileSAM, YOLOv8, and a combination of the two.

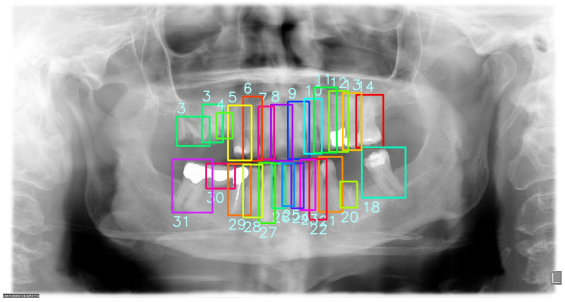


(a) Radiograph before post processing.

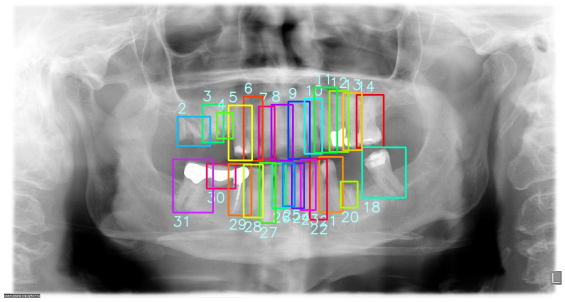


(b) Radiograph after post processing.

Figure 4. Example of a radiograph with mirrored classifications before and after post processing.



(a) Radiograph before post processing.



(b) Radiograph after post processing.

Figure 5. Example of a radiograph with a duplicate tooth 3 classification before and after post processing.

Segment Anything Model (SAM) is one of the leading segmentation models. SAM performs particularly well in zero-shot scenarios and was trained on a wide variety of datasets [6]. However, its large size and computational demands limit its practicality in low-resource environments.

MobileSAM adapts SAM to work effectively in constrained conditions [13]. MobileSAM is 50 times smaller than SAM and about 40 times faster, making it the ideal candidate for our preferably low resource model.

MobileSAM, however, has the noticeable limitation of not being able to classify separate teeth on its own. That, and the fact that MobileSAM tends to perform poorly on complex shapes with fine details, such as a group of tightly packed teeth, encouraged us to compare MobileSAM with other segmentation methods such as YOLOv8.

The benefit of YOLOv8 is that it can be trained to segment and classify individual teeth. YOLOv8 also can be trained to generate bounding boxes, which can then be fed to MobileSAM to fine tune the segmentation masks of each tooth individually. We were curious to see if this method could outperform YOLOv8's built in segmentation. These three approaches formed the basis of our segmentation experiments.

Since we are using two datasets, the way the ground truth masks were created vary slightly. The "Teeth Segmentation on dental X-ray images" dataset from Humans In The Loop uses tight polygons while the "Tufts Dental Database" uses slightly larger and more complex composite masks. Due to this difference, we tested our three approaches on each dataset individually as well as combined.

3.4. Classification

YOLOv8 not only segments, but classifies images as well. YOLOv8 was trained to detect 32 objects corresponding to each tooth using the Universal Numbering System, as in the dataset. For classification and segmentation inference, we used Class Agnostic Non-Maximum Suppression with an IoU Threshold of 0.45. This means if any bounding boxes (whether the same class or not) overlap with an IoU of 0.45 or greater, the bounding box with the lower confidence is removed.

After segmenting and classifying the data, we applied a post-processing step to enforce anatomical plausibility. This script identified and corrected predictions that would be infeasible in real-world dental anatomy, such as duplicate instances of the same tooth or teeth detected on the incorrect side of the mouth (e.g., tooth 13 predicted on the left side rather than the right). In cases where YOLOv8 mirrored a tooth to its counterpart (Fig.4a), we flipped the classification to the anatomically correct label (Fig.4b). For duplicate classifications, we confirmed adjacency and adjusted labels incrementally based on the surrounding teeth

Segmentation Method	Metric		
	Dice	IoU	Recall
Humans In The Loop Dataset			
Fine Tuned MobileSAM	0.8939	0.8083	0.8512
YOLOv8	**0.9121**	**0.8388**	**0.9336**
YOLOv8 + FT MobileSAM	0.8396	0.724	0.8111
Tufts Dataset			
Fine Tuned MobileSAM	0.9147	0.8431	0.8959
YOLOv8	**0.9358**	**0.8801**	**0.9485**
YOLOv8 + FT MobileSAM	0.8667	0.7667	0.845
Combined Dataset			
Fine Tuned MobileSAM	0.9152	0.844	0.8972
YOLOv8	**0.9277**	**0.8661**	**0.942**
YOLOv8 + FT MobileSAM	0.7629	0.6504	0.7348

Figure 6. Table showcasing results from the nine experiments performed.

(Fig. 5).

Post processing improves the quality of cropped inputs for diffusion model training and increases the frequency of correct tooth representations during model learning.

3.5. Diffusion

We chose to use Stable Diffusion as our diffusion model for this study due to its efficient latent space implementation, which significantly reduces computational cost. Its ability for inpainting and conditioning is also advantageous, as it facilitates future extensions incorporating these methods. Our implementation utilizes a U-Net architecture for the denoising process, coupled with batch normalization to accelerate training convergence. Cropped tooth images from YOLOv8 were resized to 28×28 pixels for training within the latent space and then restored to the mean height and width of their respective tooth class during inference.

4. Results

4.1. Segmentation

For segmentation, we evaluated three approaches: (1) fine-tuned MobileSAM, (2) fine-tuned YOLOv8 segmentation, and (3) fine-tuned YOLOv8 bounding boxes with fine-tuned MobileSAM segmentation. We evaluated segmentation performance using Dice similarity coefficient (DSC), Instersection over Union (IoU), and recall, computed against ground truth masks. Figure 6 shows the results from these three experiments on each dataset individually and combined. As shown, YOLOv8's fine-tuned segmentation (Fig.7) outperforms MobileSAM with and without individual bounding boxes in all categories.

4.2. Classification

For classification, we applied our post-processing script to reassign severely misplaced predictions. Prior to post-processing, YOLOv8 achieved an accuracy of 0.951 (Fig. 8). The confusion matrix revealed mirrored misclassifications (e.g., tooth 3 predicted as tooth 14), visible as northeast diagonals in the top-left and bottom-right quadrants. These errors arise from YOLOv8 internally flipping images during inference.

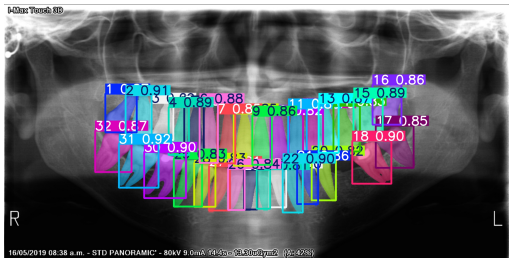


Figure 7. Example of a radiograph segmented by YOLOv8.

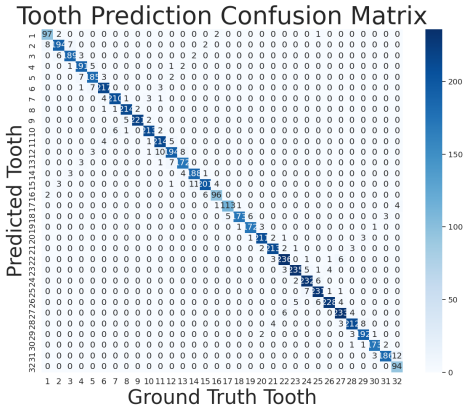


Figure 8. Example of the confusion matrix after YOLOv8 inference without any post processing. This classification had an accuracy of 0.951.

After post-processing, accuracy improved to 0.967 (Fig. 9), with the mirrored diagonals eliminated. This correction removed implausible predictions not addressed by YOLOv8 alone, enhancing the anatomical plausibility of classifications and improving downstream diffusion training by providing more accurate tooth-specific inputs. The remaining errors were primarily adjacent misclassifications (off-by-one errors), reflecting inherent difficulty in distinguishing neighboring teeth.

4.3. Diffusion

Our diffusion model, trained on cropped tooth images from the previous steps, successfully generated realistic and anatomically plausible reconstructions of missing teeth. As shown in Fig. 10, the model produces diverse outputs for the same tooth (e.g., teeth 3 and 18) while maintaining the correct morphology for the given tooth. This diversity reflects the model's ability to capture natural variation in tooth

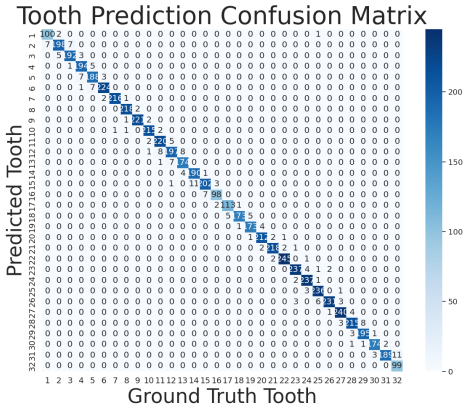


Figure 9. Example of the confusion matrix after post processing. This classification had an accuracy of 0.967.

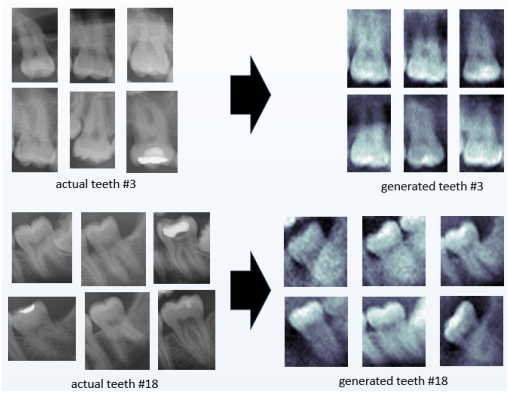


Figure 10. Example of diffusion results from generating tooth 3 and 18, showing diversity in results.

shape while preserving anatomical consistency.

These results demonstrate the potential of diffusion-based methods for tooth synthesis in dental radiographs. However, occasional imperfections were observed in the generation of thinner teeth, due to their height to width ratio. While our current evaluation is qualitative, future work will include quantitative metrics such as structural similarity (SSIM) and expert assessment to further validate clinical applicability.

5. Conclusion and Future Work

5.1. Conclusion

In this work, we developed an integrated pipeline for automated analysis and reconstruction of teeth in dental radiographs. By evaluating MobileSAM, YOLOv8 segmentation, and a hybrid YOLOv8+MobileSAM approach, we demonstrated that YOLOv8 provides superior segmentation performance across multiple metrics and datasets. Post-processing further enhanced classification accuracy by correcting anatomically implausible predictions such as mirrored and duplicate teeth. Building on these improvements, we trained a diffusion model capable of generating realistic and diverse synthetic teeth, demonstrating the feasibility of generative modeling for reconstructing missing teeth in dental radiographs. Collectively, these contributions illustrate a viable path toward automated dental charting and radiograph augmentation, offering potential benefits for clinical workflows, education, and the development of robust dental AI systems.

5.2. Future Work

Future directions include incorporating quantitative evaluation of the diffusion model using image similarity metrics (e.g., SSIM) and expert clinical assessment to validate anatomical realism. Expanding the dataset to include more diverse patient populations, varied imaging modalities (e.g., panoramic and periapical radiographs), and pathologic cases will improve model generalizability. Additionally, integrating conditional diffusion with treatment planning (e.g., generating reconstructions with implants or restorations) could directly support clinical decision-making. Finally, deploying the pipeline in real-world dental settings and conducting prospective studies will be critical to assess its clinical utility and refine the system for practical adoption in resource-rich and resource-limited environments alike.

6. Acknowledgments

We thank Dr. Mubarak Shah and Dr. Niels da Vitoria Lobo at the UCF Center for Research in Computer Vision as well as Dr. Brian Kirkwood of USAISR for their inputs on this paper. We also thank NSF REU for this opportunity.

References

[1] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. 2

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 2

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. 2

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 2

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3

[7] Humans In The Loop. Teeth segmentation on dental x-ray images, 2023. 2

[8] Vicente Reinaldo Fretes López, Carlos G. Adorno, Julio César Mello Román, José Luis Vázquez Noguera, Ricardo Gariba Silva, Horacio Legal-Ayala, Jorge Daniel Mello-Román, Ricardo Daniel Escobar Torres, and Jacques Facon. Panoramic radiography database, Jan. 2021. 2

[9] Karen Panetta, Rahul Rajendran, Aruna Ramesh, and Shishir Rao. Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 10 2021. 2

[10] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 2

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 2

[12] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. 2

[13] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 3