Advancing SAM for Dental Imaging: A Detection-Prompted Pipeline for High-Accuracy Tooth Segmentation

Saya Atchibay SAYOKIT@BU.EDU
Aiya Alchaar AIYAJE@BU.EDU
Farshid Alizadeh-Shabdiz ALIZADEH@BU.EDU
Metropolitan College, Department of Computer Science, Boston University, MA, USA

Peixi Liao LIAOPX@BU.EDU
Henry M. Goldman School of Dental Medicine, Boston University, MA, USA

Editors: Under Review for MIDL 2026

Abstract

Since the Segment Anything Model (SAM) was introduced in 2023 (Kirillov et al., 2023) numerous studies have investigated its performance on medical imaging datasets. SAM has been trained on the SA-1B dataset, which comprises 11M natural images and 1.1B segmentation masks, which makes it a very powerful engine. However, due to the differences between natural images and medical X-ray images, SAM does not perform as well as much smaller CNN based models for segmentation of medical images. To improve segmentation of dental panoramic radiographs, we propose a detection-guided prompting pipeline in which YOLOv11 localizes tooth regions and SAM generates initial masks. Although these SAM masks capture the general tooth structure, they often miss fine morphological details. To address this, we introduce a lightweight U-Net refinement module that operates purely as a post-processing step, correcting local boundary errors without fine-tuning SAM. Trained on our combined dataset, YOLOv11 achieved an mAP@0.5 of 0.9931. Our full pipeline improves SAM's zero-shot Dice score from 0.672 to 0.9026, demonstrating that detection-guided prompting coupled with lightweight refinement substantially enhances segmentation quality.

Keywords: Dental Image Segmentation, Object Detection, X-Ray images, Transfer learning, Foundation Model, SAM, YOLOv11, U-Net

1. Introduction

X-rays are an important tool in dental care. They help your dentist look at parts of the mouth that cannot be seen during a regular examination. Accurate automated analysis of dental x-ray images has various demands like in treatment planning, diagnosis, implant planning and in forensic, which would be a lot easier to determine by having a helper tool which produces high accuracy masks. There are different types of dental x-rays (Mark, 2024), and the most common ones are: bite-wings show the back teeth from the chewing surface to the bone near the gumline, periapical shows whole tooth from the crown to the tip of the root and panoramic x-rays that shows the entire jaw and all of the teeth in 1 image. In this work, we use panoramic X-ray images as our primary data source, since they capture the full dentition in a single scan and are widely used in clinical practice for comprehensive assessment.

Powerful prompt-driven foundation model Segment Anything Model (SAM) proved its high performance of segmentation on different domains. Roy et al. (2023) evaluates the zero-shot effectiveness of the SAM for medical image segmentation using few clicks and bounding box promptings demonstrating high accuracy on novel medical image segmentation tasks. Single positive bounding boxes are seen to perform considerably better than 10 positive point prompts. Khalili et al. (2024) used different prompts for SAM (bounding boxes, points and texts) to compare the effect on Lung segmentation results and it was found that the most effective approach was using sets of points obtained from the average images of each dataset. But since mouth structure varies a lot from person to person and getting average points of teeth is not feasible, in this paper we used box prompts.

Transfer learning is a common method in modern computer vision. It means using models that were trained on very large image datasets and then applying them to a new task with fewer images(Zhuang et al., 2021). Models like YOLO and SAM already learned many general visual features such as edges, textures, and basic shapes. Because of this, they can still understand new images fairly well, even if they have never seen that specific dataset before. This is especially helpful in medical imaging, where datasets are often small. With transfer learning we can get better results, faster training, and stronger performance compared to training a model from scratch.

In this work, we used two datasets to implement our approach: one is an open source UFBA-UESC Dental Dataset to train YOLOv11, and a private dataset that the Dental department at Boston University annotated with segmenting masks. Final results were evaluated using our private dataset since the public dataset has specific segmenting masks as sharp edges of the tooth whereas private dataset gives more accurate natural tooth shaped segmenting masks. We conducted an ablation study to determine which type of bounding-box prompt enables SAM to produce the most accurate segmentation masks. We compared two prompting strategies: using one bounding box per tooth versus a single box enclosing the entire dental region. In addition, we performed a threshold analysis to determine the YOLO confidence level that produces the most accurate bounding boxes.

In the following paragraph we will talk about related works that were done on improving the segmentation of x-ray images.

2. Related works

A variety of approaches have been proposed for dental X-ray segmentation and detection, ranging from classical convolutional architectures to more recent transformer-based foundation models. Prior works differ in whether they focus on detecting tooth regions, segmenting individual teeth, or enhancing segmentation performance through architectural modifications. Below, we summarize several representative methods that have informed the development of our pipeline.

Suryani et al. (2021) proposed an automatic detection system for panoramic dental X-rays using Mask R-CNN with a ResNet-101 backbone. Their model identifies candidate tooth regions through a Region Proposal Network and then predicts class labels, bounding boxes, and segmentation masks, achieving high confidence in detecting restoration areas.

Bashar et al. (2024) evaluated tooth segmentation on panoramic X-rays using six U-Net variants, each tested with either two or three convolutional layers per block. Their results

showed that most architectures produced similar-quality masks. The two-layer Vanilla U-Net was the fastest while still achieving about 88% Dice. With three layers, Dense U-Net and R2 U-Net achieved the highest Dice scores (90%), though at the cost of greater complexity and computation time.

Yang et al. (2025) proposed architectural modifications to improve SAM's performance on medical images. They augment SAM's ViT encoder with a parallel CNN branch for local texture and boundary cues and redesign the mask decoder with a U-Net–style multi-scale fusion module. This hybrid approach enables better modeling of fine anatomical structures and yields consistently higher Dice scores and boundary accuracy across multiple medical imaging datasets compared to vanilla SAM.

Xing et al. (2024) presented a full pipeline for segmenting teeth and surrounding oral structures on panoramic X-rays and using those segmentations for dental implant planning. They trained a ResUNet-based CNN on panoramic and tangential radiographs to segment teeth, bone, sinuses, restorations, nerves, and root canals, followed by post-processing to separate connected teeth and extract anatomical boundaries. Using PCA and linear regression, the system estimated implant axes that closely matched clinical plans, achieving 0.9785 pixel accuracy, 0.8483 IoU, and sub-millimeter error in implant position.

While these studies demonstrate strong progress in dental image analysis, most either rely solely on CNN-based models or require substantial architectural modifications and training complexity. In contrast, our work combines the strengths of a lightweight detector, a prompt-driven foundation model, and a minimal refinement module. By using YOLO to guide SAM and applying a small U-Net as a post-processor, our approach achieves high segmentation accuracy without the cost of end-to-end training or model redesign.

3. Datasets

This study uses two-dimensional panoramic dental radiographs capturing both teeth and jawbone structures. Two datasets were used: a private dataset (Dataset-Y) and a public UFBA-UESC subset (Dataset-M), both annotated by professional dentists.

Dataset-Y (Private): Contains 224 panoramic images with one union tooth mask (.png) per X-ray, where all teeth are merged into a single segmentation. 51 images were manually annotated by drawing bounding boxes around each tooth using the ground-truth masks.

Dataset-M (Public): Contains 425 panoramic images with individual per-tooth masks (.ome.tiff), where each tooth is segmented separately. Bounding boxes for each tooth were generated directly from their individual ground-truth (.ome.tiff) masks and then they were combined into a single PNG mask for each image so the dataset follows the same structure as our private dataset.

Images in both datasets include a wide range of variations: X-rays from different age groups, images with both low and high contrast, images with and without braces, and images containing restorations, implants, or missing teeth. They also differ in size and in how much of the tooth region is visible in each X-ray. These variations show the diversity of the dataset and support the fairness of the evaluation.

4. Research methodology

4.1. Dataset preparation

A total of 51 images with bounding boxes from Dataset-Y and 425 images from Dataset-M were combined and then split into training, validation, and testing sets (333 for training, 71 for validation, and 72 for testing). The split was done using stratification, ensuring that each set contained the same proportion of images from Dataset-Y and Dataset-M. This combined dataset was used to train the YOLOv11 model.

Only Dataset-Y was used for segmentation experiments. Dataset-M was used for training YOLOv11 because its masks use a simplified, straight-edged style that does not match natural tooth contours. Including these masks in the segmentation evaluation would distort accuracy measurements, therefore, Dataset-M was excluded from all segmentation experiments.

4.2. Background

In this section, we briefly review the main components relevant to our method: the Segment Anything Model (SAM), the YOLOv11 object detector, and the U-Net architecture, which we employ as a lightweight refinement module for improving SAM's outputs

SAM-1. The model's prompt-based design lets it handle completely new image distributions and tasks in a zero-shot manner. Kirillov et al. (2023) evaluates SAM across numerous tasks and show that its zero-shot performance is often competitive with, or even exceeds, prior fully supervised methods. Although a newer SAM-2 model exists, its main improvements target video and dynamic inputs, which are not needed in our work. Sengupta et al. (2025) shows that SAM-2 performs worse than SAM-1 on medical images with low contrast, such as CT and X-ray.

YOLOv11. Among real-time state-of-the-art object detectors, the YOLO family is known for combining high speed with strong accuracy. YOLOv11 brings several improvements compared to earlier YOLO versions. It uses an updated backbone and feature-fusion layers, which help the model extract stronger visual features for more accurate detection. The architecture is also optimized for efficiency, allowing the model to run faster while keeping a good balance between speed and accuracy (Jocher et al., 2023).

U-Net refiner. U-Net is a convolutional encoder–decoder architecture widely used in medical image segmentation due to its ability to capture both global structure and fine boundary details. Its skip connections preserve high-resolution spatial information, enabling the model to reconstruct smooth and anatomically consistent masks even in low-contrast regions.

4.3. Proposed Pipeline

4.3.1. Baseline Segmentation Using SAM

For the baseline, SAM was applied using its Automatic Mask Generator, which produces region proposals without any prompts. This configuration evaluates SAM's zero-shot performance on dental panoramic images. To isolate tooth-specific regions, we applied a post-filtering step that removed proposals that were too small or large, touched image borders, or

were located outside the main tooth area. The remaining high-confidence proposals—ranked by SAM's predicted mask quality score—were merged into a single binary mask per image.

4.3.2. Evaluation of different prompting strategies

We tested two types of bounding-box prompts for SAM. The first setup used a single bounding box per X-ray that encloses the entire set of teeth. The second setup used one bounding box per tooth, resulting in up to 32 boxes per image.This study evaluated how the granularity of bounding-box prompts influences SAM's segmentation performance. As illustrated in Figure 1, SAM behaves noticeably differently when guided by a single coarse bounding box versus multiple fine-grained per-tooth boxes.

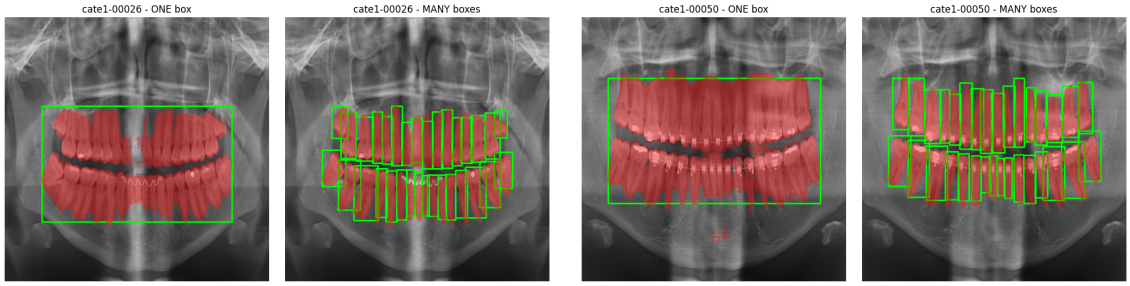

Figure 1: Two X-ray examples showing SAM segmentation with one box (left) and many boxes (right)

4.3.3. Training YOLOv11

We trained YOLOv11 using the combined and stratified dataset described in the Dataset Preparation section. The model was initialized from the pretrained YOLOv11-nano checkpoint (yolo11n.pt), which provides a lightweight starting point for transfer learning. The model was trained for 150 epochs using the default image size of 640×640 and a batch size of 16. This setup allowed YOLOv11 to adapt efficiently to the dental X-ray domain with minimal architectural changes.

4.3.4. Segmentation Using SAM With YOLO-Derived Box Prompts

For segmentation, we used the original SAM ViT-H model in inference mode without any fine-tuning. Each panoramic X-ray was first enhanced using CLAHE and converted to RGB, then passed to SAM through the SamPredictor interface. The predictor automatically performs the preprocessing required by SAM, including resizing the image so that the longest side is 1024 pixels and applying the model's internal normalization. Bounding boxes predicted by YOLOv11 were used directly as box prompts. These boxes were transformed to SAM's internal coordinate space, and SAM produced multiple candidate masks per box. For each tooth box, we selected the mask with the highest score, and the union of these masks formed the final segmentation for the image.

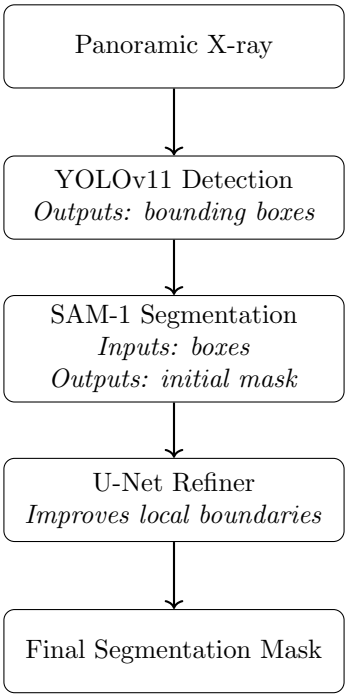

Figure 2: Methodology schema of our proposed segmentation pipeline.

Although the YOLO+SAM pipeline already produced strong results, we observed several consistent failure modes in the SAM masks (as it is shown on Figure 4). SAM often struggled to follow the true tooth shape, especially in regions with restorations or fillings and in areas with lower contrast, also tending to merge the tooth roots into a single region rather than separating the roots. In these cases, SAM tended to ignore faint or irregular parts of the tooth and produced masks that were slightly under-segmented or had missing pieces. To keep the strengths of YOLO (accurate localization) and SAM (strong zero-shot segmentation), but still correct these local errors, we added a small U-Net refiner that takes both the X-ray image and the SAM mask as input and learns to recover the missing tooth regions and sharpen the boundaries, leading to higher overall segmentation accuracy.

4.3.5. Post-processing: U-Net refiner

As a post-processing step, we trained a lightweight U-Net model to refine the YOLO-prompted SAM masks. For each panoramic X-ray, we first run YOLO+SAM on the CLAHE-enhanced image and store the union of all SAM tooth proposals. During training, the refiner takes a two-channel input: the CLAHE grayscale image and its corresponding SAM union mask, both resized to 512 x 512. The network is a small U-Net with three downsampling/upsampling stages (base width 32 channels) and skip connections, trained to predict a refined tooth mask using a combined binary cross-entropy + Dice loss. At test time, the U-Net output is upsampled back to the original image resolution. To suppress hallucinations far from SAM's initial prediction, we form a narrow band around the SAM mask by dilating it with a 21 x 21 kernel and then zero out any U-Net predictions

outside this band. The final segmentation used for evaluation is therefore the U-Net mask constrained to lie within this SAM-guided band around the teeth.

5. Experiments and Results

5.1. Baseline Segmentation results

To understand SAM's behavior without prompting, we examined its zero-shot segmentation outputs on the panoramic dataset. This baseline serves as a foundation for comparing how each subsequent component in our pipeline improves segmentation quality. Example masks from this setting are presented in Figure 3

Figure 3: Examples of zero-shot SAM-1 segmentation output.

The resulting baseline masks showed substantial variability across images. SAM often segmented only a limited portion of the dentition and frequently introduced scattered false-positive regions. These inconsistencies reflect the difficulty of applying SAM directly to panoramic X-rays without external guidance. To complement the qualitative examples shown above, we report the corresponding quantitative performance of the zero-shot SAM-1 baseline in Table 1.

Table 1: Zero-shot SAM-1 baseline segmentation results.

Dice	**IoU**	**Precision**	**Recall**
0.6720	0.5230	0.7217	0.6813

5.2. Evaluation of different prompting strategies

Table 2 shows that prompting SAM with one large bounding box per image leads to noticeably weaker segmentation performance. Using a single box results to lower Dice (0.76) and IoU (0.62), mainly because the model struggles to separate individual teeth within a large region. In contrast, providing one bounding box per tooth substantially improves all metrics. The per-tooth prompts yield much higher Dice (0.90) and IoU (0.83), as well as better precision, indicating cleaner mask boundaries with fewer false positives. Recall remains high in both settings, but the per-tooth prompts achieve a significantly more balanced precision–recall. Overall, SAM performs far better when guided with detailed, localized prompts rather than one coarse box covering the entire dental region.

Table 2: Accuracy metrics comparing SAM segmentation using one bounding box per image versus multiple per-tooth bounding boxes.

Method	**Dice**	**IoU**	**Precision**	**Recall**
SAM (big box per image)	0.7615	0.6242	0.6862	0.8911
SAM (each tooth separately)	0.9045	0.8265	0.9134	0.8977

5.3. Training YOLOv11

We trained YOLOv11 on both raw and CLAHE-enhanced images to evaluate whether contrast enhancement improves detection quality. As shown in Table 3, the CLAHE-trained model achieves a slightly higher mAP50 (0.9931 vs. 0.9915), while the remaining metrics remain almost identical across both settings. These results indicate that YOLO performs robustly regardless of preprocessing, with CLAHE providing only marginal improvements.

Table 3: Detection performance of YOLOv11 on the test set using raw images versus CLAHE-enhanced images.

Method	**mAP50**	**mAP50–95**	**Precision**	**Recall**
YOLO boxes on test set – RAW	0.9915	0.7601	0.9829	0.9870
YOLO boxes on test set – CLAHE	0.9931	0.7598	0.9823	0.9876

After training the detector, we performed an evaluation to determine the best confidence threshold for selecting YOLO boxes before passing them to SAM. We evaluated thresholds from 0.20 to 0.60 and measured the resulting segmentation accuracy on the Dataset-Y. The

Table 4: Evaluation of segmentation performance under varying YOLO confidence thresholds.

Confidence	**Dice**	**IoU**	**Precision**	**Recall**
0.20	0.8819	0.7909	0.9201	0.8525
0.25	**0.8821**	**0.7912**	**0.9211**	**0.8521**
0.30	0.8819	0.7909	0.9212	0.8518
0.35	0.8815	0.7904	0.9214	0.8511
0.40	0.8813	0.7901	0.9217	0.8506
0.45	0.8811	0.7898	0.9219	0.8499
0.50	0.8809	0.7895	0.9221	0.8494
0.55	0.8808	0.7893	0.9222	0.8490
0.60	0.8799	0.7881	0.9229	0.8473

results in Table 4 show that Dice and IoU remain stable across thresholds between 0.20 and 0.55, with a gradual decline at higher thresholds due to missing detections. Precision

slightly increases as the threshold becomes stricter, while Recall decreases as fewer boxes are kept. Based on this trade-off, we selected a confidence threshold of 0.25 for our final pipeline.

5.4. YOLO+SAM segmentation results

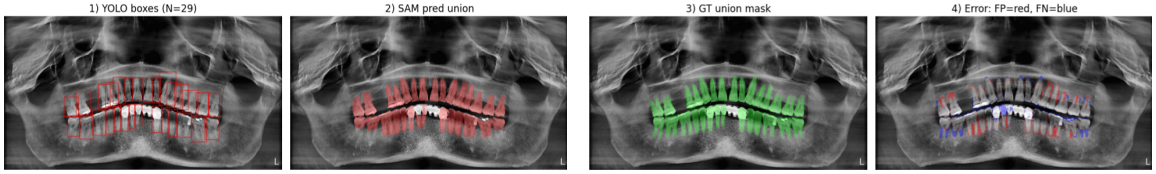

Figure 4: Example of segmentation produced by the YOLO+SAM pipeline. (1) Bounding boxes generated by YOLO. (2) SAM's predicted segmentation masks. (3) Ground-truth union mask. (4) Error visualization showing false positives in red and false negatives in blue.

As seen in Figure 4, the YOLO+SAM pipeline captures most teeth well, but the predicted masks still show several local inconsistencies. In particular, some boundaries appear softened or slightly irregular, and a few teeth exhibit small missing regions. The error map highlights these issues through scattered false positives and false negatives around the enamel contours. These observations indicate that, while SAM benefits from accurate bounding boxes, its outputs remain imperfect and require refinement to achieve smoother and more anatomically consistent masks.

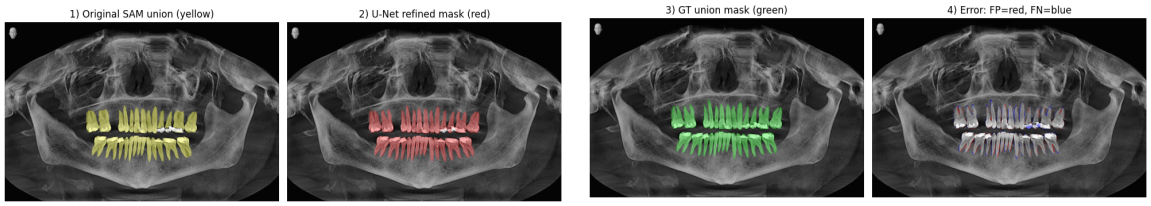

Figure 5: Example of segmentation produced by the YOLO+SAM+U-Net pipeline. (1) SAM segmentation from YOLO-prompted masks. (2) Output of the proposed U-Net refiner. (3) Ground-truth union mask. (4) Error visualization showing false positives in red and false negatives in blue.

Figure 5 illustrates how the U-Net refiner corrects several common failure modes of SAM. In the SAM-only output (1), the model incorrectly segments high-contrast restorations as individual teeth because it lacks prior knowledge of tooth morphology. After refinement (2), the U-Net restores the full tooth shapes, filling in missing regions while suppressing isolated restoration artifacts. Another limitation of SAM is that it tends to miss the separation between a tooth's roots and shows them as one connected shape. The refined mask shows that

the U-Net learns this structural pattern and segments the root contours more accurately, producing results that align much more closely with the ground-truth mask (bottom-left).

The experiments show (Table 5) a clear progression in performance: SAM alone struggles to capture full tooth structures. Introducing YOLO bounding boxes provides more precise and localized prompts, enabling SAM to extract more complete masks. Adding a small U-Net refiner corrects fine-detail errors that SAM misses. Taken together, these components form a complementary pipeline that significantly boosts segmentation accuracy, achieving the best Dice, IoU, and Recall scores in our study.

Table 5: Comparison of segmentation performance across all evaluated methods.

Method	**Dice**	**IoU**	**Precision**	**Recall**
Baseline (zero-shot SAM-1)	0.6720	0.5230	0.7217	0.6813
YOLOv11 + SAM-1	0.8821	0.7912	0.9211	0.8521
YOLOv11-CLAHE + SAM-1-CLAHE	0.8837	0.7935	0.9080	0.8662
YOLOv11-CLAHE + SAM-1-CLAHE + U -Net	0.9026	0.8231	0.9085	0.9007

5.5. Limitations

Despite the strong quantitative improvements, the proposed pipeline has several limitations. First, the approach remains sensitive to the quality of YOLO detections, as inaccuracies in bounding boxes are inherited by both SAM and the U-Net refiner. Second, the refinement stage is restricted to a SAM-guided spatial band and therefore cannot reconstruct teeth that SAM does not detect. Finally, the current framework produces a single union mask rather than separating individual teeth, which restricts applicability for tasks such as tooth indexing or pathology localization.

5.6. Conclusion

In this paper, we presented a detection-prompted segmentation pipeline that leverages YOLOv11 for tooth localization, SAM for initial mask generation, and a lightweight U-Net for refinement. Our experiments demonstrate that accurate spatial prompts are essential for guiding SAM on panoramic X-rays, and that a small refiner is sufficient to correct local structural errors without retraining SAM. The final pipeline achieves substantial improvements over all baselines, confirming the effectiveness of combining object detection with prompt-driven segmentation for dental imaging.

In future work, we aim to extend the pipeline to instance-level tooth segmentation, explore more robust prompting strategies for SAM, and test the method on larger and more diverse panoramic datasets to assess generalizability.

References

Mubtasim Bashar, Rahil Mushfiq, Amitabha Chakrabarty, Shahriar Hossain, and Yong Jung. Semantic segmentation on panoramic x-ray images using u-net architectures. *IEEE Access*, PP:1–1, 01 2024. doi: 10.1109/ACCESS.2024.3380027.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Ebrahim Khalili, Blanca Priego-Torres, Antonio León-Jiménez, and Daniel Sanchez-Morillo. Automatic lung segmentation in chest x-ray images using sam with prompts from yolo. *IEEE Access*, 12:122805–122819, 2024. doi: 10.1109/ACCESS.2024.3454188.

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. URL https://arxiv.org/abs/2304.02643.

Anita M Mark. The value of dental x-rays. *The Journal of the American Dental Association*, 155(4):356, April 2024.

Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model, 2023. URL https://arxiv.org/abs/2304.05396.

Sourya Sengupta, Satrajit Chakrabarty, and Ravi Soni. Is sam 2 better than sam in medical image segmentation? In Olivier Colliot and Jhimli Mitra, editors, *Medical Imaging 2025: Image Processing*, page 97. SPIE, April 2025. doi: 10.1117/12.3047370. URL http://dx.doi.org/10.1117/12.3047370.

D Suryani, M N Shoumi, and R Wakhidah. Object detection on dental x-ray images using deep learning method. *IOP Conference Series: Materials Science and Engineering*, 1073 (1):012058, feb 2021. doi: 10.1088/1757-899X/1073/1/012058. URL https://doi.org/10.1088/1757-899X/1073/1/012058.

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, 2024. doi: 10.1109/MIPR62202.2024.00046.

Sihan Yang, Jiadong Feng, Xuande Mi, Haixia Bi, Hai Zhang, and Jian Sun. Improved baselines with synchronized encoding for universal medical image segmentation, 2025. URL https://arxiv.org/abs/2408.09886.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.