An Efficient Approach to Scaling the Panoramic X-ray Tooth Instance Segmentation Dataset

Xukai Liu[1], Chenglong Ma[1], Shenxiao Mei[2]

Chohotech Ltd., China, Hangzhou

Abstract. This paper presents an efficient approach to scaling the Panoramic X-ray Tooth Instance Segmentation Dataset, developed for the STS 2024 Challenge/Competition. The task involves segmenting individual teeth from panoramic X-ray images, a critical problem in dental diagnostics. Given the large volume of unlabelled data in the provided dataset, we implemented a novel labeling strategy that leverages a small labeled training set combined with the powerful Segment Anything Model (SAM). By utilizing SAM's capability for generalization and transfer learning, we generated high-quality labels for the unlabelled images. This semi-automated labeling pipeline significantly improved model performance and led to the top result on the competition leader-board, demonstrating its effectiveness in large-scale medical imaging tasks.

Keywords: Active Learning, Semi-Supervision, Instance Segmentation

1 Introduction

Panoramic X-ray imaging of the human mouth is a crucial tool in dental diagnostics, providing comprehensive views of the dental structures, including individual teeth, jaws, and surrounding tissue. However, accurately segmenting each tooth from these complex X-ray images is a challenging task due to the overlapping anatomy, variations in tooth shape, and inconsistent image quality. Instance segmentation, which aims to identify each tooth as a separate instance, is particularly difficult in this context. Traditional segmentation models struggle to achieve high accuracy, especially when only small amounts of labeled data are available, which is often the case in medical datasets. This challenge, posed by the STS 2024 competition, requires developing an efficient, scalable approach to handle large volumes of unlabelled panoramic X-ray images while delivering accurate tooth instance segmentation.

Instance segmentation tasks in medical imaging often face data scarcity issues, where obtaining high-quality labeled datasets is expensive and time-consuming. Semi-supervised learning approaches, which leverage both labeled and unlabeled data, have emerged as a promising solution to this problem. State-of-the-art methods in semi-supervised segmentation often combine small amounts of labeled data with models pre-trained on large datasets. Techniques like consistency regularization, pseudo-labeling, and self-training have shown effectiveness in other medical imaging domains. In particular, recent advances in pre-trained models such as the Segment Anything Model (SAM)[2] have demonstrated impressive generalization across different segmentation tasks. SAM's ability to handle a wide range of image domains with minimal retraining makes it a strong candidate for applications in medical imaging, including dental X-rays. Despite these advancements, applying such methods to complex instance segmentation tasks, like tooth identification in panoramic X-rays, remains underexplored.

Motivated by the need for a scalable solution in the STS 2024 competition, we propose a novel approach that combines the power of SAM with semi-supervised learning to efficiently label a large dataset of unlabelled panoramic X-ray images. Our method starts with a small labeled training set and uses SAM's generalization capabilities to generate high-quality pseudo-labels for the unlabelled portion of the dataset. This significantly expands the training data without requiring extensive manual annotation, thus allowing the model to improve its segmentation accuracy. By implementing this strategy, we were able to outperform existing methods and achieve the best results on the competition leaderboard, demonstrating the effectiveness of our semi-supervised labeling and segmentation approach for large-scale medical imaging datasets.

2 Method

2.1 Prepossessing

We normalized the panoramic X-ray images to reduce the impact of brightness and contrast differences between images on model training. Specifically, we applied the following normalization steps: First, the pixel values of the input images were scaled to the range [0, 1]. Then, the images were standardized using the mean and standard deviation. This process utilized the global mean and standard deviation of the training set to ensure consistency in image features.

2.2 Proposed Method

To address the semi-supervised challenge presented by the dataset, we propose a robust automated labeling strategy aimed at maximizing the use of the unlabelled data while efficiently leveraging the limited available labeled samples. Initially, we train a baseline object detection model using the labeled data, allowing it to capture fundamental patterns and features relevant to caries segmentation and dental disease detection. This serves as the starting point for the labeling process. From this model, we then select its most confident predictions—representing the best-performing detection results—which are subsequently refined and labeled using the Segment Anything Model (SAM). SAM provides a sophisticated segmentation mechanism that enhances the precision of the object detection output by segmenting the relevant regions of interest.

To further ensure the accuracy and reliability of these SAM-labeled segments, we introduce an external oracle as an additional verification layer. The oracle evaluates and calibrates the SAM-generated results, correcting any inaccuracies and ensuring high-quality labels before they are integrated into the training process. These refined and verified labels are then fed back into the original model, which is retrained with this newly expanded dataset, thus incorporating the corrected labels into its learning process. This iterative procedure—alternating between model training, SAM-based segmentation, oracle calibration, and retraining enables the model to progressively improve with each cycle, increasing its capacity to generalize effectively across both the labeled and unlabelled data.

This iterative feedback loop not only enhances the model's performance but also progressively increases the size of the labeled dataset, as more unlabelled samples are confidently labeled with each iteration. Over time, this approach helps transition from a semi-supervised to a more fully supervised learning environment, allowing the model to achieve greater accuracy and robustness. This process, illustrated in Fig. 1.

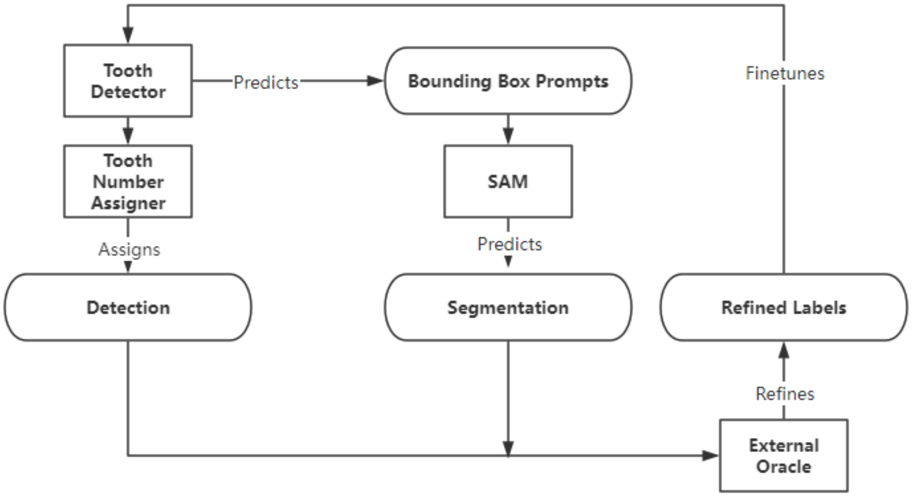

Fig. 1. Pipleline of The Data Scaling Workflow

2.3 Model Structure

For the instance segmentation task, we used the **Ultralytics YOLOv8**[1]model as our baseline. YOLOv8 is widely used in object detection tasks due to its efficiency and accuracy. It achieves an excellent balance between speed and accuracy, enabling real-time operation across various hardware environments, including applications in image and video analysis, autonomous driving, surveillance systems, and medical imaging. With our own additions, it will be more than sufficient for this task.

Improving the Proto Layer for Finer Feature Maps:

One of the core enhancements we implemented was improving the **proto layer**, which is responsible for generating the mask feature maps used in segmentation. YOLOv8, by default, provides relatively coarse feature maps, which may struggle to capture the fine-grained details necessary for accurate segmentation in dental X-rays, where tooth boundaries can be subtle. To address this, we applied additional **upsampling operations** within its segmentation portion, this serves to increase the spatial resolution of the feature maps, allowing the network to capture finer details in the images, such as the precise contours of teeth and the boundaries between overlapping dental structures. The feature maps were thus enhanced resulting in more precise segmentation masks.

Adjusting the Loss Function for Larger Bounding Boxes:

In addition to architectural improvements, we made adjustments to the **loss function**. Dental X-ray images often contain large bounding boxes that encompass entire regions of teeth, and accurately segmenting these regions requires special attention to the size and scale of the boxes. The standard YOLOv8 loss function, which is well-suited for natural images with smaller and more varied object sizes, needed to be adapted to better handle the large bounding boxes characteristic of panoramic X-rays. To achieve this, we modified the loss function to place **greater emphasis on larger bounding boxes**, ensuring that the network accurately segments entire teeth, even in cases where individual teeth are closely packed together or partially occluded. This modification improves the model's ability to distinguish between neighboring teeth and provides better segmentation boundaries.

Loss function: Our specific loss function combines several prospects[3,5] of the model output.

1. **Bounding Box Loss**: Measures the error between predicted bounding boxes and ground truth boxes using IoU loss.
2. **Objectness Loss**: Evaluates how well the model predicts whether an object is present in the bounding box, i.e. confidence of prediction.
3. **Class Loss**: Measures the classification accuracy of the predicted object class within each bounding box.
4. **Mask Loss**: Specific to instance segmentation, the loss between the predicted segmentation mask and the gt.

2.4 Post-processing

Tooth numbering presents a unique challenge, as each tooth has a specific position and structure that must be identified. Traditional object detection struggles to handle this due to potential duplicate class predictions. Instead of constraining the model to enforce unique tooth classes, we allow it to predict multiple instances per class, then resolve these conflicts in post-processing.

We prioritize predictions based on the model's **confidence scores**, selecting the highest-confidence predictions for each tooth. To finalize the assignments, we use a **Hungarian algorithm**, which efficiently matches predicted instances to tooth numbers, ensuring unique assignments based on optimal confidence scores. This method balances spatial and structural information, resolving conflicts while maintaining accurate tooth numbering.

3 Experiments

3.1 Dataset

The dataset[6] provided for the STS competition consists of panoramic X-ray images of human teeth, a category of medical imaging that presents unique challenges in terms of image complexity and detailed anatomical structures. These panoramic images capture the full dental arch, which includes teeth, jawbones, and surrounding tissue. The dataset's high variability in terms of tooth shapes, alignments, and X-ray quality makes it particularly suited for developing robust instance segmentation models.

The dataset is divided into training, validation, and testing sets as follows: Training set: Comprises 2,400 panoramic X-ray images, with only 20 images fully labeled for tooth instance segmentation and the remaining 2,380 images unlabeled. The key challenge posed by this dataset is its semi-supervised nature. With only a small fraction of the training data labeled (20 out of 2,400 images), the task is to effectively leverage the large volume of unlabeled data to improve model performance. This requires developing techniques that can generalize well from the limited labeled data, such as employing semi-supervised learning approaches. The large amount of unlabeled data adds complexity to the segmentation task, making it necessary to utilize both labeled and unlabeled examples efficiently for accurate and scalable segmentation in medical imaging applications.

3.2 Evaluation metrics

For evaluation metrics, check out the Challenge's official online introductory page, which clearly gives the evaluation metrics, as well as the final score calculation. The address of the competition is: STS 2024 Challenge

3.3 Implementation details

Environment settings For example: The development environments and requirements are presented in Table 1. The system is running Ubuntu 20.04.5 LTS as the operating system. The CPU in use is an e.g., Intel(R) Core(TM) i9-7900X CPU@3.30GHz with a clock speed of 3.30GHz. The system has a total of 64GB RAM, divided into 4 modules of 16GB each, operating at a speed of 2.67MT/s. The system is equipped with 1 NVIDIA 3090 24G GPU. The CUDA version installed on the system is 11.8. The programming language used for development is Python 3.8. The deep learning framework employed includes torch 2.4.1. These specifications provide insight into the hardware and software setup used for the development of a specific project or application.

Table 1. Development environments and requirements.

System	Ubuntu 20.04.5 LTS
CPU	Intel(R) Core(TM) i9-7900X CPU@3.30GHz
RAM	16×4GB; 2.67MT/s
GPU (number and type)	1 NVIDIA 3090 24G GPU
CUDA version	11.8
Programming language	Python 3.8
Deep learning framework	torch 2.4.1

Table 2. Training protocols.

Batch size	10
Total epochs	100
Optimizer	SGD
Initial learning rate (lr)	1e-3
Training time	4 hours
Loss function	CIoU, DFL[3], VFL[5]
Number of model parameters	
Number of flops	

4 Results and discussion

The results we produced were promising, utilizing the power of SAM, we were able to tackle the limited labelled data and generate large amount of high quality labelled from a few "seed data" points.

5 Conclusion

Our method has been able to excel in the competition of STS2024 Panoramic Teeth Segmentation, and secure a solid first place.

Acknowledgements. The authors of this paper declare that the segmentation method they implemented for participation in the STS 2024 challenge has not used any additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention. We thank all the data owners for making the X-ray images and CT scans publicly available and Codebench [4] for hosting the challenge platform.

Table 3. Quantitative evaluation results. **The results should correspond to your final algorithm submission. The internal validation denotes the performance on the internally partitioned validation cases with ground truth. The online validation denotes the leaderboard results. The Testing results will be released during MICCAI. Please leave them blank at present.** You can use a similar Table format to present the ablation study results of the public and online validation. A useful online tool to create latex table `https://www.tablesgenerator.com/latex_tables.`

Method	image-level			instance-level			
	Dice	IoU	NSD	Dice	IoU	NSD	IA
ChohoTech	0.8544	0.9218	0.8744	0.9591	0.7644	0.8555	0.8914
jichangkai	0.7982	0.9402	0.8414	0.9674	0.7566	0.8877	0.8199
camerart2024	0.7755	0.8957	0.8205	0.9312	0.6902	0.8119	0.7936
isjinghao	0.8388	0.879	0.7795	0.916	0.6845	0.7891	0.782
Your baseline model	0.7023	0.8299	0.7812	0.8023	0.5913	0.5122	0.7936
Your final model	0.8451	0.9175	0.8724	0.9560	0.7650	0.8487	0.9186

References

1. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLO (Jan 2023), `https://github.com/ultralytics/ultralytics` 3
2. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), `https://arxiv.org/abs/2304.02643` 2
3. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection (2020), `https://arxiv.org/abs/2006.04388` 4, 6
4. Xu, Z., Escalera, S., Pavao, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7) (2022) 6
5. Zhang, H., Wang, Y., Dayoub, F., Sünderhauf, N.: Varifocalnet: An iou-aware dense object detector (2021), `https://arxiv.org/abs/2008.13367` 4, 6
6. Zhang, Y., Ye, F., Chen, L., et al.: Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data **10**(1), 380 (2023) 5