Coarse-to-Fine Pseudo Annotations for Semi-supervised Teeth Segmentation

Xu Shi[1], Shengyin Yang[2], Jiacheng Wang[3], and Wentai Hou[4]

[1] The Affiliated Stomatological Hospital of Kunming Medical University, Yunnan Provincial Stomatology Hospital, Kunming 650032, China
[2] The First Dental Clinic of The Affiliated Stomatological Hospital of Kunming Medical University, Yunnan Provincial Stomatology Hospital, Kunming 650231, China
[3] Manteia Technologies, Co., Ltd., Xiamen 361005, China
[4] The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming 650118, China
jiachengw@manteia.com;houwentai@fudan.edu.cn

Abstract. Teeth instance segmentation plays a crucial role in dental diagnostics and treatment planning. However, obtaining fully annotated data for training deep learning models is time-consuming and costly. In this work, we propose a semi-supervised approach to teeth instance segmentation that leverages a small amount of labeled data alongside a larger, unlabeled dataset, which is called Coarse-to-Fine Pseudo annotations (CFP). In general, CFP follows the branch of using pretrained models to produce high-quality pseudo annotations and combining the real and pseudo annotations to re-train the model. In difference, observing that the binary segmentation performance is adequately good, CFP produces the pseudo annotations only at the stage of teeth instance classification. Our method achieved an average Dice score of 90.40% and an average NSD score of 92.74% for the teeth segmentation on the validation set using a NVIDIA GeForce RTX 4090 Ti. The average running time was 13 seconds for one image. The code will be available at http://github.com/jcwang123.

Keywords: Teeth Instance Segmentation · Pseudo Annotation.

1 Introduction

Computer-aided diagnosis tools are becoming increasingly integral in modern dental practice, particularly for tasks like treatment planning and comprehensive prognosis evaluation. Among these tools, 2D panoramic X-ray imaging and 3D cone-beam computed tomography (CBCT) are effective methods for identifying conditions like invisible caries, impacted teeth, and supernumerary teeth in children. However, one of the main challenges is that manually identifying and annotating teeth in these images is time-consuming and labor-intensive. This results in a limited number of labeled datasets, which hampers the development

of accurate deep-learning algorithms for automatic teeth segmentation and disease analysis. Semi-supervised learning has emerged as a promising alternative, as it can extract valuable information from unlabeled data to overcome this limitation. Despite its potential, significant challenges remain in optimizing these algorithms due to the scarcity of labeled cases and the complexity of dental image structures.

In the context of dental image segmentation, two primary challenges arise. First, wrong classified instances often occur due to the variability in tooth shapes, overlapping structures, and varying image quality, leading to inaccurate predictions by the model. Second, closed contexts, where teeth are closely packed or occluded by neighboring structures, make it difficult for the model to accurately differentiate individual teeth. These challenges complicate the segmentation task and hinder the overall performance of automated diagnostic systems, and at the meanwhile, increase the difficulty of pseudo annotation generation.

Semi-supervised segmentation methods have gained significant attention in recent years, leveraging both labeled and unlabeled data to improve model performance [3,7,4]. Two widely adopted techniques are the teacher-student framework and pseudo-labeling. The teacher-student model is popular in semi-supervised learning, particularly in segmentation tasks. In this method, the teacher model is first trained using labeled data. Then, the teacher generates predictions (soft labels) on the unlabeled data, which are used to train the student model. The student model is optimized using both the labeled data and the teacher's predictions. To enhance the robustness of this method, consistency regularization is often applied, where the student is encouraged to produce consistent predictions under perturbations (e.g., noise or augmentations). This framework has been effective in improving segmentation performance in medical imaging, where labeled data is scarce.

We designed a two-stage segmentation network for teeth segmentation. In the first stage, the network performs coarse localization of the teeth regions, identifying the general area where teeth are present. In the second stage, the network refines this localization by performing precise segmentation of individual teeth and classifying each tooth with its corresponding ID. The entire approach is implemented based on the nnU-Net framework, leveraging its flexibility and automated configuration for optimal performance in medical image segmentation tasks. This dual-stage strategy allows for efficient and accurate identification and segmentation of teeth, even in complex dental images.

2 Method

2.1 Preprocessing

– Before model training, we perform data cleaning and statistical analysis to ensure the dataset's consistency. This includes analyzing key characteristics such as the size, shape, and grey values of the teeth regions across the dataset. Outliers or corrupted samples are removed or corrected to enhance the quality of the data, including the small teeth with pixels less than 36.

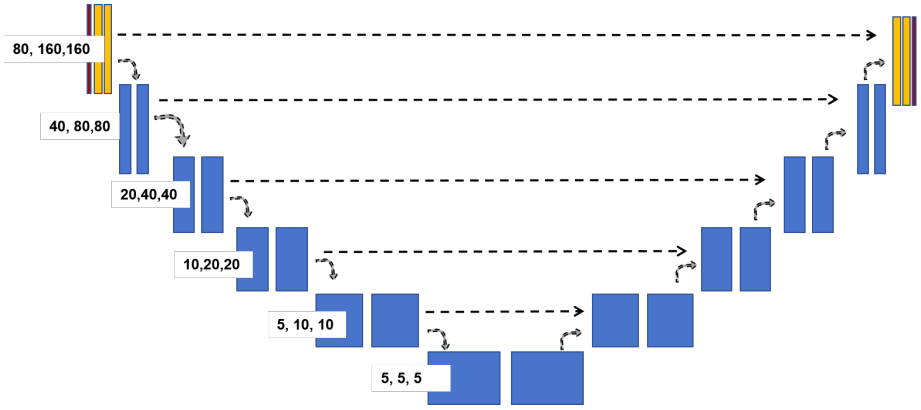

Fig. 1. Network architecture at the first stage.

- Since dental images often have anisotropic spacing, we apply resampling to normalize the voxel spacing across all images. In the first stage of our pipeline, we resample the images to a uniform spacing of 0.47 mm, which is sufficient for coarse localization. In the second stage, we use a finer resampling resolution of 0.25 mm to capture detailed tooth structures and perform accurate segmentation.
- Intensity normalization is applied to ensure consistency in image brightness and contrast. We standardize the intensity values across the dataset by applying z-score normalization, where the intensity values are normalized based on the mean and standard deviation of the dataset. This helps reduce variability due to different imaging conditions and improves model generalization.
- Additional preprocessing steps include cropping the region of interest (ROI) to focus on the dental area and reducing unnecessary background information. Augmentation techniques such as rotation, flipping, and scaling are also employed to enhance model robustness and prevent overfitting.

2.2 Proposed Method

Fig. 1 and Fig. 2 shows a typical example of 3D nnU-Net [2]. In our modified version of the 3D nnU-Net, we replace the standard convolutional blocks with residual convolutional modules. These residual blocks, first introduced in ResNet, consist of shortcut connections that bypass one or more convolutional layers. The key advantage of using residual modules is that they help mitigate the vanishing gradient problem, allowing the network to train deeper layers more effectively. This architecture change enables the model to learn more complex features while maintaining efficient gradient flow during training.

Each residual block in our modified nnU-Net contains: Two 3D convolutional layers, each followed by batch normalization and ReLU activation. A skip connection that directly adds the input to the output of the second convolution

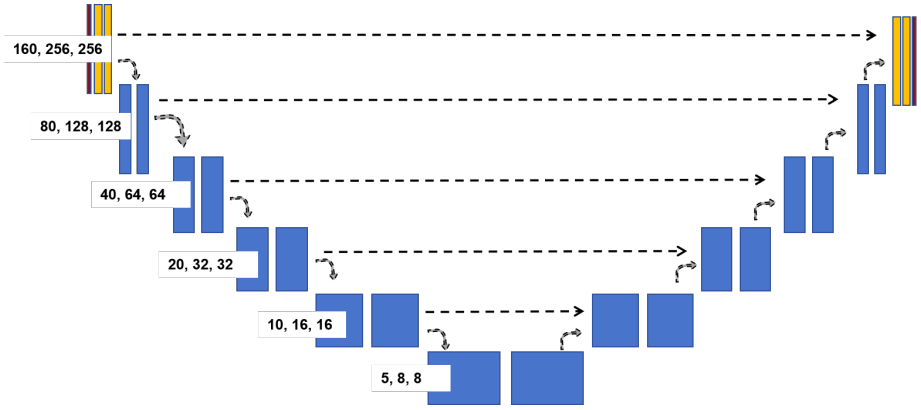

Fig. 2. Network architecture at the second stage.

layer, effectively creating a residual mapping. By incorporating these residual blocks, the network can capture more refined features with better generalization and training stability, especially for the fine segmentation tasks required in dental imaging. This modification improves segmentation accuracy while preserving the flexibility and automated configuration capabilities of the original nnU-Net framework.

We use the summation between Dice loss and cross-entropy loss because compound loss functions have been proven to be robust in various medical image segmentation tasks.

After training the second-stage model, which performs teeth segmentation, we apply this model to the unlabeled images. The predictions from this stage are treated as pseudo-labels. These pseudo-labels provide initial estimates of the tooth regions, which include rough segmentation boundaries and tooth IDs. To ensure the quality of the pseudo-labels, we perform confidence-based filtering. Predictions with high confidence scores are retained as reliable pseudo-labels, while low-confidence areas are either refined or discarded. This helps reduce the impact of noisy or incorrect predictions on the second-stage training. The second-stage model focuses on fine-grained segmentation and tooth classification. We use the high-confidence pseudo-labels from the first stage as training data for the second stage. This model performs detailed tooth segmentation and assigns tooth IDs more accurately, based on the refined input regions.

2.3 Post-processing

After segmentation, we use connected component analysis to identify and remove small, isolated regions that are likely false positives. This technique groups connected pixels or voxels into distinct components. Components smaller than a predefined size threshold are eliminated, as they are typically noise or irrelevant structures. This step ensures that only valid tooth regions remain in the final output, improving segmentation precision.

We also apply testing-time augmentation (TTA) to improve the robustness of the final segmentation. The input images are augmented using transformations like rotations, flips, and scaling, and the model predicts segmentations for each augmented version. The final segmentation output is then obtained by averaging or voting across the augmented predictions. TTA helps reduce variability due to different orientations or scales, yielding a more reliable result.

3 Experiments

3.1 Dataset

The STS dataset [2,1,6] contains 350 CBCT (Cone Beam Computed Tomography) images, which are used for the training and validation of our tooth segmentation and identification model. Labeled Data includes full manual annotations of tooth boundaries and tooth IDs, serving as the ground truth for supervised learning. Unlabeled images do not have manual annotations but are used in the semi-supervised learning process to generate pseudo-labels for further training.

3.2 Evaluation metrics

To assess the performance of our tooth segmentation and identification model, we employ several evaluation metrics. These metrics are:

- Dice Similarity Coefficient (DSC): it measures the overlap between the predicted tooth segmentation and the ground truth for each individual tooth. It is calculated as:
$$\text{DSC} = \frac{2 \times |A \cap B|}{|A| + |B|}$$
where A is the predicted segmentation and B is the ground truth segmentation. The DSC ranges from 0 to 1, with higher values indicating better segmentation quality. The DSC is also averaged across all teeth in an image to give an overall segmentation performance score for that image.
- Normalized Surface Distance (NSD) measures the distance between the predicted and ground truth surfaces for each individual tooth. It calculates how much the predicted surface deviates from the ground truth by comparing the average distance between corresponding boundary points. A lower NSD indicates a closer match to the ground truth. The NSD is also averaged across all teeth in the image, providing an overall surface accuracy score for the entire segmentation.
- Mean Intersection-over-Union (mIoU) is also known as the Jaccard index, measures the ratio of the intersection over the union of the predicted and ground truth segmentations for each individual tooth:
$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}.$$

mIoU is the average IoU across all teeth in the dataset. Higher values represent more accurate tooth segmentation. Similar to DSC, mIoU is also averaged across all teeth in a single image to provide an overall segmentation score for that image.

These metrics together provide a comprehensive evaluation of the model's performance in segmentation accuracy (DSC, NSD, mIoU), offering insights into the model's ability to accurately segment teeth and assign correct tooth IDs.

3.3 Implementation details

Environment settings The development environments and requirements are presented in Table 1. The system is running Ubuntu 18.04.5 LTS as the operating system. The CPU in use is an Intel(R) Core(R) Platinum 8358 CPU with a clock speed of 2.60GHz. The system has a total of 256GB RAM, divided into 128 modules of 2GB each. The system is equipped with six NVIDIA 4090 24G GPUs. The CUDA version installed on the system is 11.8. The programming language used for development is Python 3.60. The deep learning framework employed includes torch 2.0, torchvision 0.2.2, and monai 1.1.0. 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 18.04.5 LTS or Windows 11
CPU	Intel(R) Core(TM) Platinum 8358 CPU@2.60GHz
RAM	128×2GB;
GPU (number and type)	Six NVIDIA 4090 24G
CUDA version	11.8
Programming language	Python 3.20
Deep learning framework	torch 2.0, torchvision 0.2.2
Specific dependencies	
Code	

Training protocols Training protocols are listed below.

4 Results and discussion

Using the unlabelled cases can slightly improve the performance. The main challenges lie in the choice of the training epochs on the combined dataset. It is noticed that in cases with incomplete anatomical structures, models are more likely to predict wrong teeth IDs.

Table 2. Training protocols.

Network initialization	First Stage	Second Stage
Batch size	8	4
Patch size	80×160×160	160×256×256
Total epochs	1000	1500
Optimizer	Adam	Adam
Initial learning rate (lr)	1e-4	1e-4
Lr decay schedule		
Training time	32 hours	96 hours
Loss function		

Table 3. Training protocols for the refine model (if using two-stage framework).

Network initialization	
Batch size	2
Patch size	80×192×160
Total epochs	1000
Optimizer	SGD with nesterov momentum ($\mu = 0.99$)
Initial learning rate (lr)	0.01
Lr decay schedule	halved by 200 epochs
Training time	72.5 hours
Number of model parameters	41.22M[5]
Number of flops	59.32G[6]
CO_2eq	1 Kg[7]

4.1 Qualitative results on validation set

In case 0020 in Fig. 3, most teeth can be correctly segmented and classified.

5 Conclusion

In this paper, we have presented a semi-supervised learning method based on coarse-to-fine pseudo annotations. We test the method on a public challenge dataset. The results of the official validation set have indicated the improvements. However, most failure cases are wrongly classified and are hard to predict correctly after training on the pseudo annotations. The use of pseudo-labels mainly contributes to the improvement of segmentation metrics.

Table 4. Quantitative evaluation results.

Method	Inf. T.	image-level			instance-level			
		Dice (%)	IoU (%)	NSD (%)	Dice (%)	IoU (%)	NSD (%)	IA (%)
U-Net	-	83.45	83.45	83.45	83.45	75.56	45.21	45.56
Your baseline model	13s	95.02	90.55	97.52	88.47	83.91	90.47	95.50
Your final model	13s	96.07	92.58	98.29	90.40	87.50	92.74	95.86

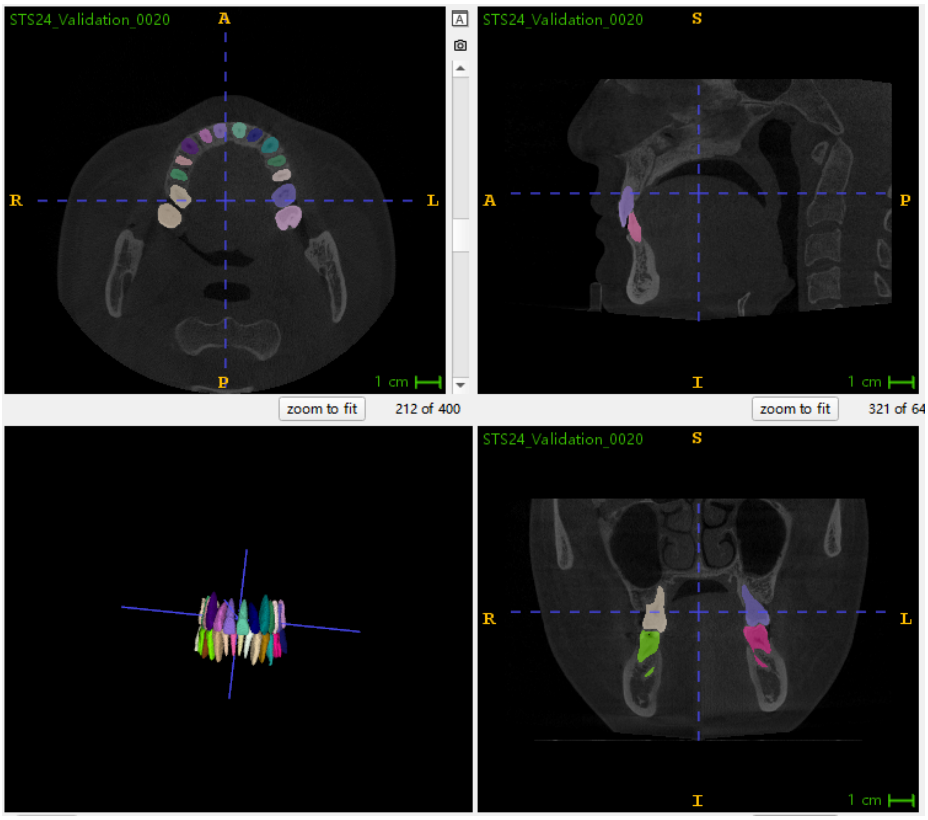

Fig. 3. Qualitative results on case 0020.

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 [5] for hosting the challenge platform.

References

1. Cui, W., Wang, Y., Zhang, Q., Zhou, H., Song, D., Zuo, X., Jia, G., Zeng, L.: Ctooth: a fully annotated 3d dataset and benchmark for tooth volume segmentation on cone beam computed tomography images. In: International Conference on Intelligent Robotics and Applications. pp. 191–200. Springer (2022) 5
2. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021) 3, 5
3. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013) 2
4. Lu, L., Yin, M., Fu, L., Yang, F.: Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. Biomedical Signal Processing and Control **79**, 104203 (2023) 2
5. Xu, Z., Zhao, H., Tu, W.W., Richard, M., Escalera, S., Guyon, I.: Codabench: Flexible, easy-to-use and reproducible benchmarking for everyone. arXiv preprint arXiv **2110** (2021) 8
6. Zhang, Y., Ye, F., Chen, L., Xu, F., Chen, X., Wu, H., Cao, M., Li, Y., Wang, Y., Huang, X.: Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data **10**(1), 380 (2023) 5
7. Zheng, X., Fu, C., Xie, H., Chen, J., Wang, X., Sham, C.W.: Uncertainty-aware deep co-training for semi-supervised medical image segmentation. Computers in Biology and Medicine **149**, 106051 (2022) 2

Table 5. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	4
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	1, 2
Pre-processing	3
Strategies to use the partial label	4
Strategies to use the unlabeled images.	4
Strategies to improve model inference	4
Post-processing	4
The dataset and evaluation metric section are presented	5
Environment setting table is provided	1
Training protocol table is provided	2
Ablation study	4
Efficiency evaluation results are provided	4
Visualized segmentation example is provided	3
Limitation and future work are presented	Yes
Reference format is consistent.	Yes