A Novel Two-Stage Approach for 3D Dental Tooth Instance Segmentation

Yuhan Chen[1], Chunshi Wang[2], and Bin Zhao[2,3](✉)

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
[2] School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
zhaobinnku@mail.nankai.edu.cn
[3] Guangxi Colleges and Universities Key Laboratory of AI Algorithm Engineering, Guilin, Guangxi, 541004, China

Abstract. Tooth CBCT instance segmentation is a foundational step in advancing digital dental health systems, with the precision of tooth segmentation playing a critical role in accurate medical diagnosis. Traditional neural networks face significant challenges in accurately locating and classifying teeth in 3D dental images, especially given the complex anatomical structures present in CBCT scans. Additionally, manually identifying and labeling each tooth from these scans is an extremely time-consuming and burdensome task for medical professionals. To address these issues, in this paper, we propose a two-stage semi-supervised method for tooth instance segmentation, along with innovative data preprocessing. Our network achieved a distinguished fourth place in the validation set of the "MICCAI STS 2024 Challenge Task 2," outperforming other mainstream semi-supervised networks.

Keywords: Tooth segmentation · Semi-supervised learning · Cone beam computed tomography

1 Introduction

As AI technology rapidly advances, computer-assisted medicine has found wide application in the dental field, especially in treatment planning and comprehensive prognosis evaluation. CBCT (Cone Beam Computed Tomography) is an advanced 3D imaging technology that is widely used in dental diagnosis and treatment planning.

In past research, researchers tended to design simple manual methods for tooth segmentation, such as relying on the contrast between teeth and surrounding tissues. However, when the data source changes, these methods often require complex manual interventions to maintain accuracy[2, 1]. With the development of convolutional neural networks (RNN), many fully automated segmentation algorithms have emerged, with nnUnet demonstrating exceptional segmentation performance. To further enhance the model's performance, many segmentation

algorithms focus on integrating tooth information into the network learning process. For example, Tae Jun Jang and others[3] use 3D images to generate panoramic images of the upper and lower jaws, thereby improving the accuracy of tooth classification. Cui and colleagues[4] extract edge maps of teeth to enhance image contrast at the shape boundaries, aiding the network's training. Additionally, Cui and others[5] explore the combination of centroid and skeleton networks to achieve effective segmentation of tooth shape and type. However, these methods often involve complex and large network structures, requiring extensive data and cumbersome processes to achieve the desired segmentation results, limiting their practicality and efficiency in real-world applications.

However, despite previous methods achieving some effect in tooth segmentation, these methods still rely on fully supervised learning and face significant challenges in obtaining annotated data. Manually identifying and annotating teeth from CBCT scans is not only time-consuming and labor-intensive but also makes it difficult to acquire large-scale annotated case data. In situations where data is lacking, advancements in the semi-supervised domain offer us new perspectives[7, 8]. How to fully utilize unlabeled data has become a key aspect in the rapid development of the semi-supervised learning field.

In the field of semi-supervised learning, consistency regularization is widely used to extract features from unlabeled data. This method assumes that the model's predictions of the same sample under different disturbances should remain consistent. For example, FixMatch[10] ensures the model produces consistent predictions for different versions of the same image by applying random disturbances to the input images, helping it learn features from unlabeled data. The Mean-Teacher model[11] achieves semi-supervised learning by splitting into a teacher model and a student model, where the pseudo-labels generated by the teacher model guide the student model's learning on unlabeled data, effectively utilizing the unlabeled data. CPS[12] uses weakly disturbed one-hot labels to supervise the outputs of strongly disturbed inputs, encouraging the network to apply consistency regularization to images with different levels of disturbances. Unimatch[13]learns features from unlabeled data by applying feature disturbances to the network.

Most 3D models use an encoder-decoder structure for feature extraction and target segmentation of CBCT images. However, we noticed that although the commonly used 3D VNet performs well in binary segmentation of teeth, it still struggles to distinguish between different tooth categories. This observation made us consider the necessity of incorporating 2D segmentation information into the decoding network. Specifically, 2D segmentation can provide more detailed boundary information for different tooth categories, which helps improve classification accuracy. Therefore, we decided to modify the original segmentation network structure to effectively integrate 2D segmentation information and enhance the model's capability in category recognition.

So, we proposed a multi-stage semi-supervised method for tooth instance segmentation, which secured fourth place on the validation set in the MICCAI2024 semi-supervised challenge.

2 Materials and Method

2.1 DataSet

The dataset comes from STS2024-3d and consists of 3D CBCT images. The training set includes a total of 330 CBCT images (30 labeled cases and 300 unlabeled cases)[6, 9]. The validation set has 20 images, and the test set is not yet available.

2.2 Method

Our model is divided into two main stages: extracting the ROI region and using semi-supervised techniques for multi-task segmentation. Our overall framework is shown in Figure 1.

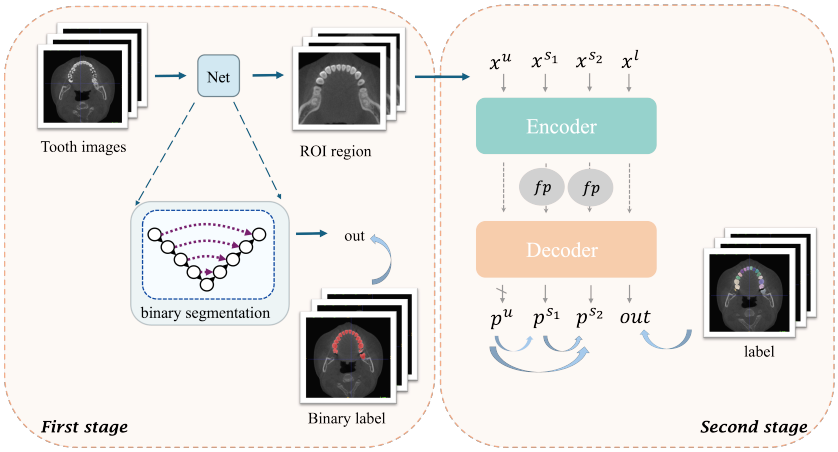

Fig. 1. Two-Stage Semi-Supervised Tooth Segmentation Framework

In this article, we define a 3D image as $\mathbf{X} \in R^{W \times H \times L}$, and the goal of semi-supervised segmentation is to obtain a label map for each voxel, defined as $\mathbf{Y} \in \{0, 1, ..., C-1\}^{W \times H \times L}$, where C represents the number of categories. The training set $\mathbf{D}$ is divided into N labeled data points $\mathbf{D}^l = X_i^l, {Y_i^l}_{i=0}^N$ and M unlabeled data points $\mathbf{D}^u = X_i^u, {Y_i^u}_{i=N+1}^{M+N}$.

In the first stage, we use a simple VNet network to learn the contour features of the teeth and extract the ROI region. This network is trained with binarized tooth images as labels (i.e., regions where $Y^l > 0$ as the foreground) through a semi-supervised binary segmentation method. This approach enhances the learning capability for both labeled and unlabeled data. After binary segmentation of the unlabeled data, we further use morphological operations, specifically opening operations, to refine the segmentation results. Opening operations combine

erosion and dilation steps, first eroding the image to remove small noise points and tiny connections, then dilating to restore the main structure of the object. Specifically, we use a 12×12 structural element for the morphological opening operation to eliminate noise interference and avoid the impact of irrelevant areas, thereby achieving a more precise region of interest (ROI).

In the second stage, we use a network with mask assistance for multi-class segmentation training. The architecture of this network is as follows:

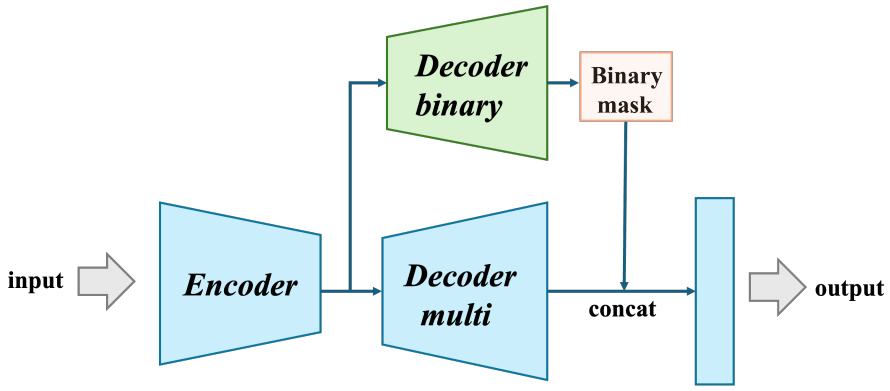

Fig. 2. The network for adding tooth contour auxiliary information to a multi-classification network.

We noticed that the network performs well when only segmenting tooth contours without classifying tooth types. So, we adjusted the network architecture to include one encoder and two decoders. The encoder extracts information from the entire 3D patch; the first decoder focuses on binary segmentation of tooth contours, and this binary segmentation result is then combined with the multiclass segmentation decoder. This design allows the multiclass segmentation decoder to classify more accurately under the guidance of the tooth contours, thereby improving the overall performance of the model.

We're using the Unimatch framework for our research. For labeled data, we employ a supervised network to generate results, while for unlabeled data, we utilize strong augmentation to handle it. In the bottleneck part of the network, we introduce feature perturbations to the strongly augmented data, using the non-augmented results to supervise the output of the strong augmentation, and making the results of the strong augmentation supervise each other. This allows the network to fully extract features from the unlabeled data.

2.3 Evaluation metrics

We evaluate the model using image-level and instance-level metrics as well as recognition accuracy. At the image-level and pixel-level, we use three metrics:

Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD), and mean Intersection-over-Union (mIoU). The formulas for DSC and mIoU are as follows,

$$Dice = \frac{2 * |G \cap P|}{|G| + |P|}, \tag{1}$$

$$mIoU = \frac{|G \cap P|}{|G \cup P|}, \tag{2}$$

In this context, G and P represent the actual segmentation and the predicted segmentation, respectively.

To calculate the normalized surface distance, you first need to find the real surface overlap area O_G and the predicted surface overlap area O_P:

$$O_G = \sum_i A_G(i) \quad (d_{G \to P}(i) \leq T), \tag{3}$$

$$O_P = \sum_j A_P(j) \quad (d_{P \to G}(j) \leq T), \tag{4}$$

$$NSD = \frac{O_G + O_P}{\sum_i A_G(i) + \sum_j A_P(j)}, \tag{5}$$

Here, T represents the tolerance, $A_G(i)$ and $A_P(j)$ denote the areas of the actual surface and the predicted surface respectively, while $d_{G \to P}(i)$ and $d_{P \to G}(j)$ stand for the distances from the actual surface to the predicted surface and from the predicted surface to the actual surface respectively.

For each instance category, if the current recognized mIoU > 0.5, the category is considered correctly identified. The proportion of all correctly identified categories in the total categories is counted as the identification accuracy (IA), calculated as follows:

$$IA = \frac{\sum_{c=1}^{C} mIoU}{C}, mIoU_c > 0.5 \tag{6}$$

It means C represents the number of categories, and mIoU_c indicates the mIoU for the c-th category.

3 Experiment

3.1 Data Process

To prevent precision loss caused by 3D image scaling, we didn't resample or resize the images. Instead, we used random-crop to get them to (112, 112, 80) dimensions for the network input. To make the most out of the training samples, we selected one sample from the training set to use as the validation set in our experiments.

To better study tooth features, we've set a threshold range for the teeth (500 to 2500). Within this range, the network can effectively capture tooth contour information while excluding surrounding tissue interference. This threshold choice not only improves the accuracy of feature extraction but also enhances the model's sensitivity to tooth morphology. By focusing on data within this interval, the network can more clearly identify tooth boundaries and shapes, thereby boosting instance segmentation and recognition performance, reducing noise and artifacts, and ensuring more precise segmentation results. The effect of the threshold setting is shown in Fig. 3.

We didn't use the traditional method (normalizing pixel intensity distribution within the [0.5, 99.5] range). Instead, we opted for a more efficient approach. We scaled all the data to the 0-1 range, which significantly reduces the artifacts' impact on the images. By uniformly scaling all data to the 0-1 range, we can handle artifacts in the images more effectively while ensuring the stability and consistency of the image data. This method helps minimize image distortion caused by artifacts, thereby improving the overall quality and performance of image processing.

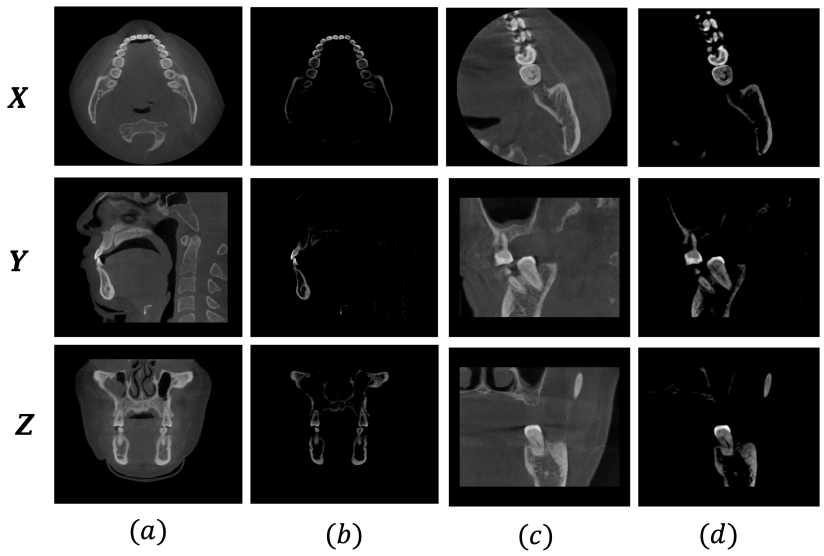

Fig. 3. The images after threshold processing, from top to bottom, are the processed images of cross-sectional, sagittal, and coronal views, respectively.

3.2 Implementation Details

In this study, we used the AdamW optimizer for model training, setting the learning rate to 0.01 and the weight decay coefficient to $w_decay = 0.0001$. During

Table 1. The top five results in the STS2024-3d competition on the validation set, with our method marked in red.A represents the instance level, while B represents the image level.

participant	Score	$Dice_A$	$Dice_B$	NSD_A	NSD_B	$mIoU_A$	$mIoU_B$	IA
Jichangkai	0.937	0.905	0.967	0.928	0.991	0.879	0.939	0.950
houwentai	0.922	0.891	0.952	0.916	0.978	0.851	0.910	0.957
ChohoTech	0.920	0.883	0.950	0.920	0.986	0.841	0.904	0.954
Guet-IICI	0.909	0.877	0.960	0.887	0.980	0.831	0.924	0.906
haoyuuuu	0.900	0.856	0.959	0.872	0.973	0.822	0.921	0.904

training, the learning rate was dynamically adjusted using the following formula:

$$lr = lr_{base} \times \left(1 - \frac{i}{N}\right)^{0.9} \tag{7}$$

In this context, lr_{base} represents the initial learning rate, which we've set to 0.01. The variable i stands for the number of iterations, and N is the maximum number of iterations, which we've set to 40,000.

The tests were conducted using a system featuring an Intel Core i7-6800K CPU, 64 GB of RAM, and an NVIDIA Tesla V100 GPU with 32 GB of memory. PyTorch was utilized for implementing all neural networks.

3.3 Results

The network we trained achieved the segmentation results shown in Figure 4 and secured fourth place in the validation set of the "MICCAI STS 2024 Challenge Task 2".

Table 2. Quantitative evaluation results of our method and comparison methods on the validation set.

participant	Score	$Dice_A$	$Dice_B$	NSD_A	NSD_B	$mIoU_A$	$mIoU_B$	IA
Mean-Teacher[11]	0.751	0.793	0.836	0.843	0.794	0.692	0.774	0.760
FixMatch[10]	0.793	0.776	0.823	0.784	0.851	0.762	0.801	0.804
CPS[12]	0.857	0.828	0.895	0.901	0.813	0.804	0.878	0.855
Ours	0.909	0.877	0.960	0.887	0.980	0.831	0.924	0.906

When using traditional 3D networks for segmentation, the tooth shape features shown in Figure (d) are hard to recognize. However, after introducing auxiliary tooth contour information, the tooth contours become much clearer. This shape information not only boosts feature recognition capability but also makes it possible to classify teeth into different categories.

We took three samples from the training set to use as a validation set and trained the network using 27 labeled samples and 300 unlabeled samples. Then,

we compared our network to other state-of-the-art networks in semi-supervised learning, such as FixMatch, Mean Teacher, and CPS. The experimental results showed that our network had the best performance.

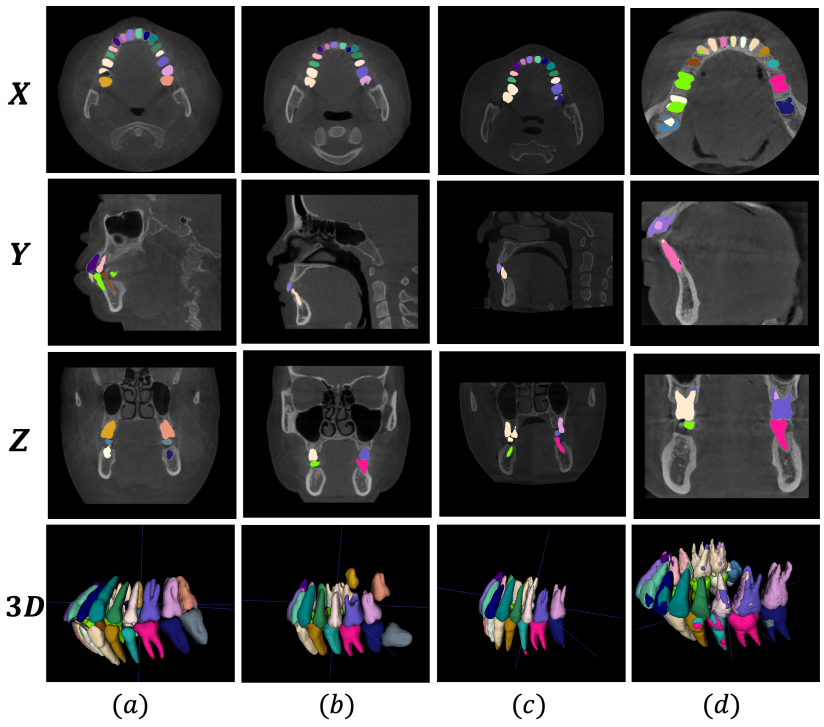

Fig. 4. Our method's segmentation results, from top to bottom, are cross-sectional, sagittal, coronal, and 3D display.

4 Conclusion

In this paper, we present a novel two-stage semi-supervised method for tooth segmentation. This method first innovatively preprocesses the data to extract the ROI (Region of Interest) area and then adds auxiliary tooth contour information during the multi-category segmentation process. Compared to other semi-supervised methods, our experiments have validated the effectiveness of our network.

Acknowledgments. This work is supported in part by the Project of Improving the Basic Scientific Research Ability of Young and Middle-Aged Teachers in Uni-

versities of Guangxi Province (Grant No.2023KY0223), Youth Science Foundation of Guangxi Natural Science Foundation (Grant No.2023GXNSFBA026018) and the Guangxi Science and Technology Major Project (Grant No.AA22068057), National col- lege students'innovation and entrepreneurship training program(Grant No.202410595074).

References

1. Jin, L. J., et al. "Global burden of oral diseases: emerging concepts, management and interplay with systemic health." Oral diseases 22.7 (2016): 609-619.
2. Gao, Hui, and Oksam Chae. Individual tooth segmentation from CT images using level set method with shape and intensity prior. Pattern Recognition 43.7 (2010): 2406-2417.
3. Tae Jun Jang, Kang Cheol Kim, Hyun Cheol Cho, and Jin Keun Seo. A fully automated method for 3d individual tooth identification and segmentation in dental cbct. IEEE transactions on pattern analysis and machine intelligence, 44(10):65626568, 2021.
4. Zhiming Cui, Changjian Li, and Wenping Wang. Toothnet: automatic tooth instance segmentation and identification from cone beam ct images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6368–6377, 2019.
5. Zhiming Cui, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang, Yuhang Sun, Lei Ma, Jiawei Huang, et al. A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. Nature communications, 13(1):2096, 2022.
6. Weiwei Cui, Yaqi Wang, Qianni Zhang, Huiyu Zhou, Dan Song, Xingyong Zuo, Gangyong Jia, and Liaoyuan Zeng. 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, pages 191–200, 2022.
7. Bin Zhao, Shuxue Ding, Hong Wu, Guohua Liu, Chen Cao, Song Jin, and Zhiyang Liu. Automatic acute ischemic stroke lesion segmentation using semi-supervised learning. International Journal of Computational Intelligence Systems, 14(1):723733, 2021.
8. Bin Zhao, Zhiyang Liu, Guohua Liu, Mengran Wu, Chen Cao, Song Jin, Hong Wu, and Shuxue Ding. Combine unlabeled with labeled mr images to measure acute ischemic stroke lesion by stepwise learning. IET Image Processing, 16(14):39653976, 2022.
9. Weiwei Cui, Yaqi Wang, Yilong Li, Dan Song, Xingyong Zuo, Jiaojiao Wang, Yifan Zhang, Huiyu Zhou, Bung san Chong, Liaoyuan Zeng, et al. Ctooth+: A largescale dental cone beam computed tomography dataset and benchmark for tooth volume segmentation. In MICCAI Workshop on Data Augmentation, Labelling, and Imperfections, pages 64–73, 2022.
10. Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33:596–608, 2020.
11. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30, 2017.

12. Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26132622, 2021.
13. Yang, Lihe, et al. "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.