Two-Stage Semi-Supervised nnU-Net Framework for Tooth Segmentation in CBCT Images

Changkai Ji[1 [0009−0007−7090−7360]], Yusheng Liu[1 [0009−0004−2624−9223]], Lanshan He[1 [0009−0009−6093−8803]], Yuxian Jiang[1 [0009−0002−7689−5333]], Chuanyi Huang[1 [0009−0009−3223−0082]], and Lisheng Wang[1 [0000−0003−3234−7511]]

Institute of Image Processing and Pattern Recognition, Department of Automation,
Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
{changkaiji, lswang}@sjtu.edu.cn

Abstract. Automated tooth segmentation in Cone-Beam Computed Tomography (CBCT) images is crucial for various dental applications, including treatment planning and computer-assisted dental prosthesis design. The MICCAI STS 2024 Challenge Task 2 aims to enhance automated tooth segmentation by providing a dataset comprising both labeled and unlabeled CBCT images. This paper addresses the challenge of limited labeled data by reformulating the problem as a semi-supervised learning task. We propose a two-stage deep learning model based on nnU-Net. Our approach initially employs a low-resolution nnU-Net for quadrant segmentation, followed by a full-resolution nnU-Net for fine tooth segmentation within each quadrant. To efficiently utilize unlabeled data, we implement a selective stability-based retraining strategy to generate reliable pseudo-labels. Our method is quantitatively evaluated on the STS 2024 validation set, achieving good performance across various metrics (Dice_instance = 90.74%, Dice_image = 97.70%). The proposed approach achieved one of the highest rankings in the competition's validation phase, demonstrating its efficacy in automatically segmenting teeth from CBCT images.

Keywords: CBCT Tooth Segmentation · Semi-supervised Learning · nnU-Net

1 Introduction

The efficacy of digital dental technology in orthodontics and implantology hinges on precise tooth segmentation and accurate labeling in Cone Beam Computed Tomography (CBCT) images, as well as efficient algorithm execution [2]. This technology not only furnish practitioners with accurate three-dimensional tooth models and correct International Dental Federation (FDI) numbering but also process data within clinically acceptable timeframes to facilitate the swift development of personalized treatment plans. Nevertheless, current technological solutions encounter numerous challenges in achieving these objectives.

The morphological similarities among teeth, particularly between adjacent and symmetrical ones, significantly complicate accurate segmentation [1]. These resemblances challenge algorithms in precisely distinguishing and delineating individual teeth, especially in cases of high-density dental arrangements. Moreover, quality issues in CBCT images further compound this challenge [5]. Artifacts, noise, and insufficient resolution can compromise the accuracy of segmentation and FDI numbering [9]. Such image quality deficiencies not only increase the complexity of segmentation and labeling processes but may also lead to algorithmic misclassification or omission of crucial anatomical details.

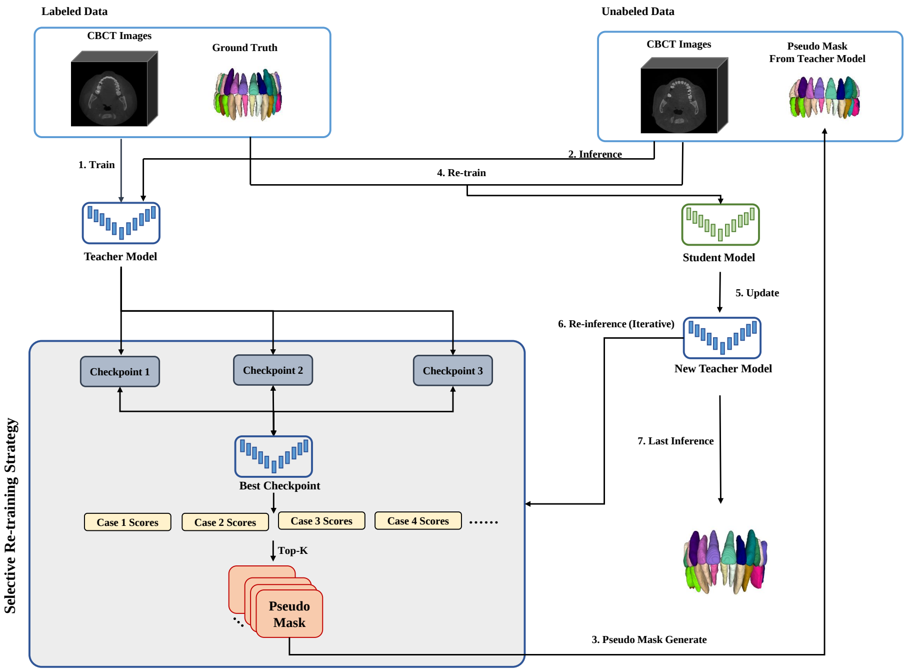

Fig. 1. The framework comprises a teacher model and a student model. The teacher model, initially trained on a limited set of labeled data, generates pseudo-labels for unlabeled data. These pseudo-labels undergo a selective retraining strategy to filter out low-quality labels. Subsequently, the student model is trained using both the original labeled data and the filtered pseudo-labeled data. In each iteration, the student model is then promoted to become the new teacher model.

Existing tooth segmentation methodologies can be categorized into two primary groups [6]: semi-automatic approaches based on traditional knowledge and automatic techniques leveraging deep learning. Traditional knowledge-based methods typically rely on manually designed features that are segmented and numbered using predetermined rules and algorithms. While effective for normal dental structures, these methods reveal limitations when confronted with complex patient conditions, struggling to adapt to lesions or malocclusions and accurately assign FDI numbers. Furthermore, these approaches often necessitate manual intervention, impeding full automation and hindering the optimization of clinical workflows.

Conversely, deep learning methods, utilizing techniques such as Fully Convolutional Networks (FCNs), automatically extract features from data without human intervention [3]. These approaches demonstrate significant potential for automation, exhibiting particularly robust performance on large-scale datasets and theoretically capable of learning both segmentation and FDI number labeling tasks concurrently. However, a notable limitation of deep learning methods is their tendency to disregard anatomical relationships between teeth, especially the natural connections between adjacent and symmetrical teeth [7]. This oversight not only compromises the accuracy and consistency of segmentation but may also result in erroneous FDI number labeling, particularly when processing complex anatomical structures or low-quality images. Moreover, intricate models designed to enhance accuracy may increase computational demands, potentially affecting real-time performance.

Both traditional and deep learning methods may encounter performance challenges when processing full-resolution three-dimensional images due to large data volumes and high computational complexity [8]. In clinical settings, prolonged processing times can impede diagnostic and treatment efficiency, thereby reducing overall workflow productivity.

These challenges underscore the necessity for developing novel algorithms in dental image processing. Ideal solutions should significantly enhance processing speed while ensuring high-precision segmentation and accurate FDI numbering. This necessitates a balanced approach to algorithm design, considering accuracy, automation degree, and operational efficiency. To address these challenges in the MICCAI STS 2024 Challenge Task 2, we propose a semi-supervised model based on nnU-Net [4], aiming to achieve efficient dental instance segmentation while considering algorithmic runtime. The contributions of our work can be summarized as follows:

- We designed a nnU-Net-based self-training framework to enhance model performance through selective iterative training.
- Our algorithm improves segmentation accuracy while maintaining runtime efficiency, striving to achieve an optimal balance between precision and performance.
- At the end of the validation phase of the competition, our method attained a top position, further demonstrating its effectiveness in the task of tooth segmentation.

2 Proposed Method

2.1 Overall Architecture

As shown in Fig. 1, the proposed segmentation framework employs nnU-Net as the foundational network model for tooth segmentation, integrating it with a self-training method for semi-supervised semantic segmentation. The methodology is elucidated below.

2.2 Stability-Based Selective Retraining Strategy

We employ a selective retraining scheme to expand the labeled dataset, prioritizing the reliability of unlabeled samples [10]. The reliability or uncertainty of unlabeled images is quantified by assessing the overall stability of pseudo-masks that evolve iteratively during training. This approach enables the selection of more reliable and predictive unlabeled images based on their evolutionary stability throughout the training process.

As depicted in Fig. 1, while supervising the teacher model's training using labeled data, we preserve several model checkpoints. The discrepancies in these checkpoints' predictions for unlabeled images serve as a reliability metric. Given that trained models tend to converge and achieve optimal performance in later training stages, we evaluate the average Dice coefficient between each early pseudo-mask and the final mask. After deriving stability scores for all unlabeled images, we rank the entire unlabeled set accordingly. Subsequently, we co-supervise the student model's training using the original labeled data and the filtered pseudo-labeled data, updating the student model to a new iteration of the teacher model.

2.3 Two-Stage nnU-Net Model

To implement this semi-supervised learning framework, a robust segmentation model is essential. Considering the specificity and complexity of medical image segmentation, we select nnU-Net as the core segmentation model in the iterative process. Fig. 2 illustrates our proposed method, which automates tooth segmentation from dental CBCT images in a coarse-to-fine manner.

Our method comprises three primary steps:

- A low-resolution nnU-Net segments four quadrants of teeth from the input CBCT image, concurrently assigning a specific class label to each quadrant.
- Based on the low-resolution nnU-Net output, a full-resolution nnU-Net segmentation network annotates teeth in each quadrant. The network was trained as a nine-class segmentation model, including eight tooth classes in the current quadrant and one class for teeth in the other quadrants.
- The segmentation results from the four quadrants are merged to form a comprehensive tooth segmentation result.

This two-stage approach balances inference time efficiency while maintaining segmentation accuracy.

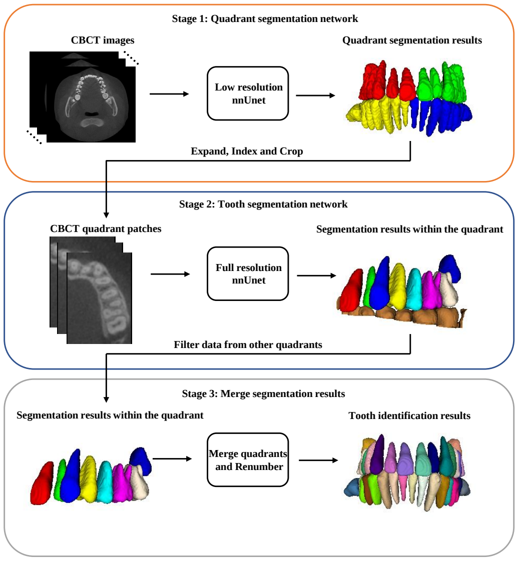
Stage 1: Quadrant segmentation network
CBCT images
Quadrant segmentation results
Low resolution nnUnet
Expand, Index and Crop
Stage 2: Tooth segmentation network
CBCT quadrant patches
Segmentation results within the quadrant
Full resolution nnUnet
Filter data from other quadrants
Stage 3: Merge segmentation results
Segmentation results within the quadrant
Merge quadrants and Renumber
Tooth identification results

Fig. 2. Schematic diagram of the two-stage segmentation method. The first stage divides all teeth into four quadrants. The second stage identifies and segments each tooth in the quadrants. The third stage combines the results from all four quadrants to reconstruct the segmentation in the original image space.

3 Experiments and Results

3.1 Dataset and Evaluation Metrics

We conducted experiments utilizing the dataset from Task 2 of the 2024 MICCAI STS Challenge. The training set comprises both labeled and unlabeled datasets. Specifically, the labeled dataset contains 30 CBCT images with precise individual tooth annotations, each accompanied by corresponding FDI tooth numbers. The unlabeled dataset consists of 300 CBCT images, while the validation set includes 20 CBCT images.

The evaluation metrics encompass both segmentation accuracy and operational efficiency measures. Accuracy metrics include instance-level and image-level Dice Similarity Coefficients (DSC), Normalized Surface Distance (NSD), Mean Intersection over Union (mIoU), and Identification Accuracy (IA). Furthermore, the challenge assesses algorithmic runtime and GPU memory consumption to comprehensively evaluate the segmentation algorithm's performance.

3.2 Implementation details

Training Set Reconstruction. To enhance the model's segmentation accuracy on small-horizon data, we select 30% of the full-volume data from the annotated training set to construct small-horizon samples during each data iteration. This approach aims to improve the model's generalization ability by generating two small-horizon samples per full-volume dataset.

Selective Re-training Strategy. We assessed image reliability using three checkpoints (at 1/3, 2/3, and 3/3 of the total rounds, respectively) uniformly saved during training. We performed two iterations of pseudo-labeling updates, with the final model trained using 30 labeled datasets, 82 pseudo-labeled datasets, and 16 small-horizon samples generated from the labeled data.

Environments and Requirements. The proposed method's environments and requirements are detailed in Table 1.

Table 1. System Configuration

Ubuntu version	Ubuntu 18.04.5 LTS
CPU	Intel(R) Core(TM) i9-10920X CPU @ 3.50 GHz
RAM	126 GB
GPU	1 NVIDIA GeForce RTX 3090 (24G)
CUDA version	12.2
Programming language	3.8.5
Deep learning framework	PyTorch (torch 1.12.1, torchvision 0.13.0)
Code will available at	After the release of the test rankings

Training Procedure. In the first stage of training the quadrant segmentation model, we disabled nnU-Net's symmetric data enhancement strategy to mitigate the influence of tooth symmetric structure on the model. When generating Regions of Interest (ROI) based on quadrant segmentation results, we applied a 10-voxel margin extension to each quadrant boundary. This approach ensures the inclusion of critical peri-dental regions, thereby facilitating more precise delineation of tooth boundaries in subsequent processing stages. During the training process for the intra-quadrant tooth segmentation model, we re-enabled nnU-Net's symmetric data augmentation strategy to fully utilize tooth structure features. To ensure that the model learnt sufficiently, we trained 300 epochs in each iteration.

Inference Acceleration. Considering that the running time of the algorithm was also evaluated during the testing phase of the competition, we optimized the computational efficiency. Since the Challenge is instance segmentation involving multiple classes, traditional interpolation methods may lead to significant computational overheads. To address this issue, we performed interpolation on floating-point numbers in tensor form using the interpolate function from PyTorch. For more details, please refer to our code. This approach not only maintains floating-point precision but also significantly improves computational efficiency when dealing with large-scale multi-class segmentation problems.

3.3 Results and Analysis

We investigated the impact of disabling mirror enhancement on segmentation performance, with results summarized in Table 2. This comparative analysis demonstrates the positive effects of this modification across various evaluation metrics.

Table 2. Comparison of Segmentation Performance with Different Configurations

	Dice (instance)	Dice (image)	NSD (instance)	NSD (image)	mIoU (instance)	mIoU (image)	IA
w/ Mirror	89.97%	96.48%	92.24%	98.48%	87.03%	93.25%	95.16%
w/o Mirror	90.74%	96.70%	93.08%	98.97%	88.00%	93.64%	95.54%

As evidenced in Table 2, disabling mirror enhancement improved all metrics. The instance-level Dice coefficient increased from 89.97% to 90.74%, while the image-level Dice coefficient rose from 96.48% to 96.70%. The NSD also improved, with instance-level NSD increasing from 92.24% to 93.08% and image-level NSD from 98.48% to 98.97%.

The mIoU metric showed similar improvements. Instance-level mIoU increased from 87.03% to 88.00%, and image-level mIoU from 93.25% to 93.64%. Additionally, the IA increased from 95.16% to 95.54%.

These results indicate that disabling mirror enhancement enhances the model's ability to accurately segment tooth structures in CBCT images. This improvement validates our hypothesis that disabling mirror enhancement would enable the model to better distinguish FDI tooth numbers. The concurrent improvement in both instance-level and image-level metrics suggests that this approach not only enhances overall segmentation quality but also improves the model's ability to recognize individual tooth instances.

Based on these findings, we opted to disable mirror image enhancement for the segmentation quadrant model in the first stage of the final segmentation framework. This version of the results was submitted as our final model for the validation stage.

Performance Comparison. We compared the two-stage nnU-Net model with the one-stage nnU-Net model. The one-stage model directly inputs the entire CBCT image into nnU-Net for instance segmentation, without employing the strategy of segmenting quadrants prior to individual teeth. As illustrated in Table 3, experimental results indicate comparable performance between the two models. However, considering the faster inference time of the two-stage nnU-Net model, we posit that it holds greater promise for clinical applications.

We believe that the main reason for the slower inference of the one-stage nnU-Net model is the large time overhead of the sliding window during inference. One potential solution involves adjusting the step size, but this approach may compromise model accuracy. An alternative strategy entails designing a sliding window technique that minimizes inference over background regions.

Table 3. Comparison of One-stage and Two-stage nnU-Net Models

	Dice	Dice	NSD	NSD	mIoU	mIoU	IA
	(instance)	(image)	(instance)	(image)	(instance)	(image)	
One-stage	90.36%	96.81%	92.52%	98.91%	87.63%	93.84%	95.57%
Two-stage	90.74%	96.70%	93.08%	98.97%	88.00%	93.64%	95.54%

Our two-stage approach mitigates GPU overhead and enhances inference speed while maintaining high accuracy by initially employing a low-resolution nnU-Net model for quadrant segmentation, followed by a full-resolution nnU-Net for tooth segmentation within the ROI. Consequently, despite the marginal performance difference, we selected the two-stage nnU-Net model for final submission.

Visualization. As shown in Fig. 3, we have visualised the segmentation results for both small-horizon data and full-volume data. The figure presents the quadrant segmentation results, the intra-quadrant tooth segmentation results, and the final results after stitching the quadrant segmentation results for both data types. The quadrant segmentation results are relatively coarse due to the low-resolution model used for quadrant segmentation to improve computational efficiency. However, this strategy effectively achieves the purpose of generating ROI while saving computational time. Subsequent full-resolution segmentation within the ROI yields more refined results, underscoring the advantages of our two-stage approach. Notably, the segmentation results for small-horizon data are comparatively inferior, highlighting the necessity of manually producing small-horizon data prior to training.

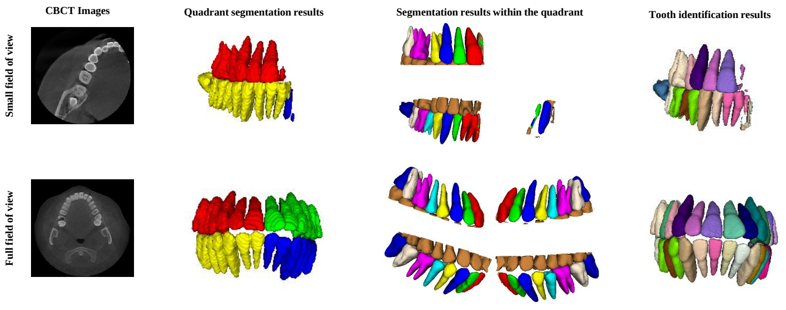
CBCT Images
Quadrant segmentation results
Segmentation results within the quadrant
Tooth identification results
Small field of view
Full field of view

Fig. 3. Visualization results of small field of view and full field of view.

4 Conclusion

In this paper, we present a two-stage semi-supervised learning framework based on nnU-Net for automatic tooth segmentation in CBCT images. Our method initially performs quadrant segmentation using a low-resolution nnU-Net model, followed by fine tooth segmentation using a full-resolution nnU-Net model within each quadrant. This strategy enhances computational efficiency while maintaining highly accurate segmentation results. We employ a stability-based selective retraining strategy to obtain reliable pseudo-labels from limited labeled data, effectively expanding the training dataset. Results demonstrate that our method achieves superior segmentation performance, achieving one of the highest rankings in the validation phase of the MICCAI STS 2024 Task 2 challenge.

References

1. Akhoondali, H., Zoroofi, R., Shirani, G.: Rapid automatic segmentation and visualization of teeth in ct-scan data. Journal of Applied Sciences **9**(11), 2031–2044 (2009)
2. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al.: A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. Nature communications **13**(1), 2096 (2022)
3. Cui, Z., Li, C., Chen, N., Wei, G., Chen, R., Zhou, Y., Shen, D., Wang, W.: Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model. Medical Image Analysis **69**, 101949 (2021)
4. 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)
5. Li, P., Liu, Y., Cui, Z., Yang, F., Zhao, Y., Lian, C., Gao, C.: Semantic graph attention with explicit anatomical association modeling for tooth segmentation from cbct images. IEEE Transactions on Medical Imaging **41**(11), 3116–3127 (2022)
6. Lian, C., Wang, L., Wu, T.H., Liu, M., Durán, F., Ko, C.C., Shen, D.: Meshsnet: Deep multi-scale mesh feature learning for end-to-end tooth labeling on 3d dental surfaces. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 837–845. Springer (2019)
7. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
8. Scholl, I., Aach, T., Deserno, T.M., Kuhlen, T.: Challenges of medical image processing. Computer science-Research and development **26**, 5–13 (2011)
9. Wu, X., Chen, H., Huang, Y., Guo, H., Qiu, T., Wang, L.: Center-sensitive and boundary-aware tooth instance segmentation and classification from cone-beam ct. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 939–942. IEEE (2020)
10. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)