CBCT Tooth and Pulp Segmentation via Watershed Fusion and Lightweight Dual-Branch nnU-Net

Rusheng Xu[1] [0009-0009-0691-718X], Zhou Yang[1] [0009-0002-3974-1183], Ting Yan[2] [0000-0001-7771-5360], and Bin Wang[1] [0000-0002-7780-0757]

[1] College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China
{xurusheng0373, yangzhou0121}@link.tyut.edu.cn, wangbin01@tyut.edu.cn
[2] Department of Pathology, Shanxi Key Laboratory of Carcinogenesis and Translational Research on Esophageal Cancer, Shanxi Medical University, Taiyuan, China
enntei@hotmail.com

Abstract. Cone-beam computed tomography (CBCT) provides essential 3D anatomical details for dental diagnosis and treatment planning. However, adjacent tooth contact, unclear pulp boundaries, and high 3D computational costs often lead to instance touching and slow processing. To address these problems, we propose an efficient segmentation method based on watershed fusion and a lightweight dual-branch nnU-Net. The high-resolution instance branch uses a marker-controlled watershed algorithm based on spatial topology to separate touching instances and refine pulp contours. Meanwhile, the semantic branch reduces input resolution (1.0 mm) to expand the receptive field, using global dental arch context for FDI tooth identification. The instance regions then act as spatial boundaries to combine semantic predictions through majority voting. Our approach was validated by our 2nd place finish in the MICCAI STSR2026 Challenge Task 1 with a final test score of 0.7308. Experimental results show that our method achieves $DSC_{\text{img}} = 0.9609$, $DSC_{\text{inst}} = 0.7525$, and IA $= 0.6668$ on validation, while significantly lowering memory requirements and improving computational efficiency, offering an effective solution for CBCT clinical applications.

Keywords: CBCT · Tooth instance segmentation · nnU-Net

1 Introduction

Cone-beam computed tomography (CBCT) provides clear three-dimensional (3D) structural information for modern dentistry, serving as a key tool for orthodontics, dental implants, and root canal treatment [8, 16]. Driven by deep learning, automated medical image analysis has made substantial progress in recent years [2, 9, 12].

Automated tooth analysis generally includes three levels: semantic segmentation (separating teeth from background), instance segmentation (isolating individual teeth), and tooth identification (assigning standard FDI numbers) [16]. Because adjacent teeth are closely arranged and share similar intensity levels, 3D tooth instance segmentation remains challenging [8]. In addition, manual 3D annotation requires heavy labor and suffers from variations near root apices and contact areas [2].

For 3D volumetric medical segmentation, modern architectures have evolved from standard 3D convolutional models like UNet++ [23] and gated attention networks like Attention U-Net [14] to vision transformers such as UNETR [6] and Swin UNETR [5]. Meanwhile, self-configuring deep learning frameworks like nnU-Net [7] have established strong baselines across biomedical segmentation tasks. In dental imaging, automated frameworks have shown good performance in multi-structure segmentation, as seen in challenges like ToothFairy2 [1]. However, standard semantic models struggle to split closely touching teeth. ToothSeg [17] attempts to resolve this via auxiliary bounding boxes, but directly training neural networks to predict instance boundaries often causes error propagation.

To solve instance touching, lower computational costs, and improve performance with limited annotations, we propose a lightweight dual-branch nnU-Net framework with watershed fusion. Evaluated on the official benchmark hidden test set, our method achieved $DSC_{\text{img}} = 0.8495$, $DSC_{\text{inst}} = 0.7122$, and $\text{IA} = 0.5976$ on the hidden test set, securing 2nd place in Task 1 of the MICCAI STSR2026 Challenge with a final score of 0.7308. Our main contributions are:

1. **Watershed Fusion**: We combine instance region predictions with spatial topology to separate closely touching teeth without adding extra network parameters.
2. **Lightweight Semantic Branch**: By setting the input voxel spacing to 1.0 mm, the semantic branch captures global dental arch context for FDI tooth numbering while greatly reducing GPU memory footprint.
3. **Optimized Inference Pipeline**: We use parallel branch execution and unified spatial resampling to shorten inference time significantly.
4. **Semi-Supervised Learning**: We design a pseudo-label filtering strategy based on anatomical rules to use unannotated CBCT data effectively when manual labels are limited.

Our implementation code and pre-trained models are publicly available on GitHub at `https://github.com/thisxu/TestMICCAI26`.

2 Method

2.1 Overall Framework

As illustrated in Fig. 1, our framework divides tooth instance segmentation and FDI tooth numbering into two collaborative branches using a dual-branch nnU-Net. This design allows the model to learn fine local structures and global dental arch relationships separately.

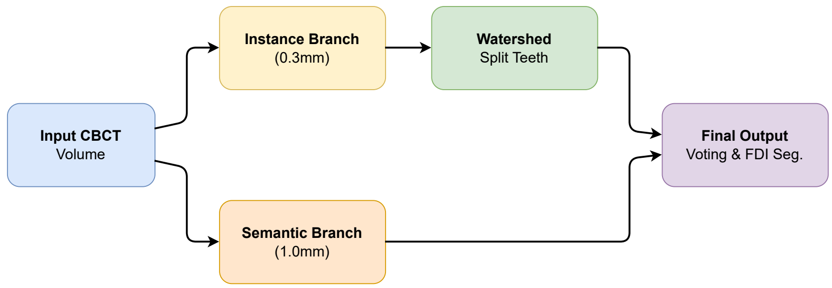

Fig. 1. Overview of the proposed lightweight dual-branch nnU-Net framework. The pipeline consists of a high-resolution instance branch for fine tooth/pulp region extraction, a coarse-resolution (1.0 mm) semantic branch for global FDI tooth identification, marker-controlled watershed separation, and instance-semantic label fusion via majority voting.

The **instance branch** uses a high-resolution 3D nnU-Net. It takes raw CBCT volumes as input and outputs three-class probability maps for background, teeth, and pulp. The predicted tooth regions are processed by a marker-controlled watershed algorithm to split touching teeth. At the same time, pulp predictions are assigned to their corresponding tooth instances to complete joint tooth, pulp, and root canal segmentation.

The **semantic branch** uses a lightweight nnU-Net operating at a coarse input resolution (1.0 mm voxel spacing). The lower resolution expands the network receptive field, allowing it to perceive overall dental arch layout, jaw relationships, and left-right symmetry for tooth identification while keeping computational costs low.

In the **fusion stage**, individual tooth instance regions obtained from the watershed separation serve as spatial limits. The semantic predictions within each instance region are tallied, and majority voting determines the final FDI number for each tooth.

2.2 nnU-Net Baseline

Our method builds upon the standard 3D full-resolution nnU-Net architecture [7], which relies on a standard 3D U-Net backbone [3, 15]. It consists of three main steps:

1. **Preprocessing**: Automated target spacing resampling, non-zero cropping, and foreground Z-score intensity normalization [7].
2. **Architecture and Loss**: A 3D U-Net backbone with deep supervision, optimized using a compound Dice and Cross-Entropy loss to tackle class imbalance.
3. **Augmentation and Training**: Extensive on-the-fly data augmentations (e.g., scaling, rotation, noise) optimized via SGD with polynomial learning rate decay [7].

4. **Inference and Post-Processing**: Gaussian-weighted sliding-window inference followed by connected-component analysis to remove background artifacts [7].

2.3 Dual-Branch Network Design

Instance Branch. The instance branch retains high spatial resolution to capture detailed structures such as crown edges, root tips, and narrow pulp canals. To avoid convergence difficulties from direct instance ID prediction, this branch predicts structural regions (background, tooth, pulp) first. These predicted region maps are then processed by the watershed algorithm to yield separated individual tooth instance regions along with their corresponding pulp and root canal sub-regions.

Lightweight Semantic Branch. FDI tooth numbering relies more on global spatial positions across the dental arch than on fine edge details. Processing high-resolution volumes directly increases memory usage and limits the receptive field. By setting the input spacing of the semantic branch to 1.0 mm, a single input patch covers a larger portion of the dental arch. Furthermore, because FDI numbering depends on left-right spatial orientation, mirror augmentation is disabled during training to prevent label confusion.

2.4 Watershed Instance Separation

To separate touching teeth without extra parameters, we apply a marker-controlled watershed algorithm [13] to the instance branch outputs. First, Euclidean distance transforms are calculated inside the predicted tooth mask. Local peaks in the distance map serve as center markers for individual teeth. The watershed algorithm then grows regions from these markers, separating touching tooth bodies cleanly.

2.5 Instance-Semantic Fusion

Instead of predicting instance IDs and class labels simultaneously within a single network, we decouple these two tasks. The instance masks obtained from watershed separation define distinct spatial regions. Even if watershed separation over-segments a single tooth into multiple regions, our region-level probability aggregation compensates for this limitation. Within each region, predicted class probabilities from the semantic branch are aggregated independently, and majority voting assigns the correct FDI label to each over-segmented fragment. This strategy ensures semantic consistency across regions and prevents prediction noise from affecting class assignment.

2.6 Semi-Supervised Training with Anatomical Filtering

To leverage unannotated data, we employ a semi-supervised training framework guided by anatomical priors. Specifically, a teacher model trained on 30 annotated CBCT scans generates pseudo-labels for 299 unannotated volumes. To eliminate noisy predictions and anatomical anomalies, these pseudo-labels are subjected to a rigorous two-level bounding-box filtering strategy:

1. **Single-Structure Constraint**: An individual component, such as a single tooth, a pulp, or a combined structure of tooth and pulp, is retained only if its 3D bounding box satisfies $\max(D, H, W) \leq 100$ voxels.
2. **Arch-Block Constraint**: Maxilla and mandible are divided into 8-tooth contiguous blocks. A block is retained only if its bounding box satisfies $W_{\text{block}} \leq 200$ voxels and $D_{\text{block}} \leq 150$ voxels.

Following this filtering, exactly 247 qualified unannotated volumes passed the criteria and were integrated with the 30 annotated scans to form the final training set.

2.7 Inference Optimization

We optimize the original nnU-Net inference workflow for clinical deployment. The instance and semantic branches run in parallel, and input volumes undergo unified spatial resampling using SimpleITK to avoid repeated interpolation. Test-time mirror augmentation is disabled, and sliding-window parameters are tuned to reduce redundant computation.

3 Experimental Setup

3.1 Dataset

Experiments were conducted on the official 3D CBCT benchmark from the MICCAI STSR2026 Challenge [18, 19]. To evaluate our method with limited annotations, the dataset was divided into:

- **Annotated Training Set**: 30 CBCT volumes with expert manual annotations for tooth regions, pulp cavities, root canals, and FDI tooth labels.
- **Unannotated Training Set**: 299 raw CBCT volumes used for semi-supervised training with pseudo-label filtering.
- **Validation Set**: 20 independent, expert-annotated CBCT scans reserved strictly for quantitative model evaluation.

3.2 Evaluation Metrics

To comprehensively evaluate our framework, the evaluation metrics are divided into two distinct categories: segmentation accuracy and computational efficiency.
Segmentation Accuracy Metrics:

– **Image-Level Dice ($DSC_{\mathbf{img}}$)**: Standard volumetric Dice overlap computed globally across the entire tooth mask to assess overall structure coverage.
– **Instance-Level Dice ($DSC_{\mathbf{inst}}$)**: Average Dice score calculated per individual tooth instance, reflecting local shape and boundary precision.
– **Instance Recognition Accuracy (IA)**: The percentage of correctly identified tooth instances with an Intersection over Union (IoU) > 0.5, evaluating joint instance segmentation and FDI numbering performance.

Efficiency Metrics:

– **Peak GPU Memory (VRAM)**: The maximum GPU memory footprint (GB) consumed during inference, reflecting hardware deployment feasibility.
– **Inference Time**: The average time required to process a 3D CBCT volume.

3.3 Experimental Environment and Protocols

All models were implemented using PyTorch and nnU-Net. The detailed software and hardware configurations used for model training, validation, and inference are summarized in Table 1. Single-volume inference speed and memory footprint were benchmarked on a single GPU.

Table 1. Hardware and software configuration of the experimental environment.

System	Ubuntu 24.04 LTS
CPU	Intel Xeon w5-3435X
RAM	256 GB (8×32 GB)
GPU (number and type)	2× NVIDIA RTX 4090 (48G)
CUDA version	12.4
Programming language	Python 3.10
Deep learning framework	PyTorch 2.5.1, nnU-Net v2.6.4

4 Results and Discussion

4.1 Quantitative Results and Ablation Study

In addition, the full training protocols, hyperparameter settings, computational resources, and computational complexity metrics required by the challenge are detailed in Table 2.

We conducted an ablation study on the validation set to quantitatively evaluate the impact of inference optimization, parallel branch execution, and the lightweight coarse-resolution semantic branch. The overall quantitative metrics, peak GPU memory footprint, and average inference runtime are detailed in Table 3.

Table 2. Training protocols and complexity.

Pre-trained Model	None
Batch size	2 per GPU
Patch size	$160 \times 128 \times 96$ (Inst.) / $128 \times 128 \times 112$ (Sem.)
Total epochs	250 (Supervised) / 1000 (Semi-Supervised)
Optimizer [7]	SGD ($m = 0.99$, wd $= 3 \times 10^{-5}$)
Initial learning rate (lr) [7]	0.01
Lr decay schedule [7]	Polynomial decay (exponent = 0.9)
Training time	$\sim$ 12 hours per branch
Loss function [7]	Compound Dice + Cross-Entropy Loss
Number of parameters	29.38 M (Inst.) / 29.36 M (Sem.)
Computational complexity	421.27 GFLOPs (Inst.) / 449.47 GFLOPs (Sem.)

Note: **Inst.**: Instance branch; **Sem.**: Coarse-resolution (1.0 mm) semantic branch.

Table 3. Ablation study of proposed optimization modules on the validation set.

Method	Inf.	Par.	Light.	$DSC_{\text{img}} \uparrow$	$DSC_{\text{inst}} \uparrow$	IA↑	VRAM (GB)↓	Time (s)↓
Baseline (0.3 mm)				0.9619	0.7532	0.6675	40.01	72.28
+ Inf.	✓			0.9609	0.7527	0.6667	40.01	26.27
Ours (Full Model)	✓	✓	✓	**0.9609**	**0.7525**	**0.6668**	**10.84**	**12.58**

Note: **Inf.**: Optimized sliding window and TTA removal; **Par.**: Parallel branch execution; **Light.**: Coarse 1.0 mm semantic branch.

As presented in Table 3, optimizing sliding-window inference and disabling unnecessary test-time augmentation (TTA) dramatically cut average per-volume inference time from 72.28 s down to 26.27 s. This achieves over a 2.7× speedup while maintaining high accuracy with $DSC_{\text{img}} = 0.9609$.

Furthermore, integrating the 1.0 mm coarse semantic branch alongside parallel branch execution reduced peak VRAM consumption from 40.01 GB to 10.84 GB, representing a 72.9% reduction. This setup further accelerated overall inference to 12.58 s, delivering a speedup of more than 5.7× over the baseline. Importantly, the impact on segmentation performance is minimal, maintaining $DSC_{\text{img}} = 0.9609$, $DSC_{\text{inst}} = 0.7525$, and IA $= 0.6668$. These results validate that global FDI tooth numbering tasks can safely run at lower spatial resolutions with negligible accuracy loss.

4.2 Qualitative Results

Figure 2 displays qualitative segmentation and 3D reconstruction outputs of our full model on Validation Case 018. Visualizations and cross-sectional inspections were performed using ITK-SNAP [21].

As shown in Fig. 2(a), our marker-controlled watershed fusion resolves interdental touching between adjacent molars, yielding precise instance boundaries. Furthermore, the high-resolution instance branch preserves fine apical and root

(a) Axial View (b) Tooth 3D (c) Pulp Perspective

Fig. 2. Qualitative results on Validation Case 018: (a) axial slice view of tooth instance segmentation, (b) 3D surface reconstruction of individual tooth instances with FDI color coding, and (c) semi-transparent perspective view of extracted pulp cavities.

canal structures, while the lightweight semantic branch ensures accurate FDI numbering across the dental arch (Fig. 2(b, c)).

4.3 Discussion

Our dual-branch architecture balances accuracy and efficiency by decoupling instance localization from global FDI numbering. The marker-controlled watershed resolves interdental touching without extra parameters, while the coarse 1.0 mm semantic branch significantly lowers VRAM usage and runtime. Compared to end-to-end models [4, 17, 22], our framework enables practical deployment on consumer-grade GPUs with less than 12 GB of VRAM and remains robust under limited annotations through semi-supervised learning.

Limitations and Future Work: Severe metal artifacts from restorations can distort image intensities and degrade watershed reliability. Future work will investigate artifact-robust methods to improve segmentation performance in restored dentitions.

5 Conclusion

We proposed a lightweight dual-branch nnU-Net framework with watershed fusion for CBCT tooth, pulp, and root canal instance segmentation. By decoupling local structural prediction and global semantic numbering into two dedicated branches, our method effectively resolves adjacent tooth touching via a marker-controlled watershed algorithm while capturing global arch context through a coarse-resolution semantic branch. Combined with semi-supervised training and parallel inference optimizations, our framework achieves high segmentation accuracy with significantly reduced GPU memory footprint and faster processing speed, offering a practical and robust solution for clinical CBCT analysis.

Acknowledgements. This work was developed and evaluated within the ODIN 2026 Challenge Cluster, which encompasses the STS, ToothFairy, and Bite2Text challenges [10, 11, 18, 19]. We thank all the data owners for making the medical images publicly available and Codabench [20] for hosting the challenge platform.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting maxillofacial structures in CBCT volumes. In: Proceedings of Optical/Medical Imaging Conferences. pp. 112–124 (2025) 2
2. Chen, H., Tian, Y., Kong, Y., et al.: Deep learning in 3D dental image analysis: A review. Computers in Biology and Medicine **148**, 105872 (2022) 1, 2
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016) 3
4. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., Liu, Y., Zhao, Y., Lian, C., Ding, Z., Zhu, M., Shen, D.: A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nature Communications **13**(1), 2096 (2022) 8
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 272–284. Springer (2022) 2
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 574–584 (2022) 2
7. Isensee, F., Jaeger, P.F., Kohl, S.A.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) 2, 3, 4, 7
8. Joda, T., Derksen, W., Wittneben, J.G., Kuehl, S.: Artificial intelligence in removable prosthodontics and dental implantology. Clinical Oral Implants Research **32**(S21), 4–15 (2021) 1, 2
9. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017) 1
10. Lumetti, L., Di Bartolomeo, M., Pellacani, A., Anesi, A., Grana, C., Bolelli, F.: Ontology-grounded structured prediction for dental cbct reporting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026) 9
11. Lumetti, L., Rizzo, F., Cremonini, F., Candeloro, E., Luca, L., Grana, C., Bolelli, F.: Do multimodal llms understand intraoral dental data? dataset, platform, and baselines. In: European Conference on Computer Vision (ECCV) (2026) 9
12. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik **29**(2), 102–127 (2019) 1
13. Meyer, F.: Topographic distance and watershed lines. Signal Processing **38**(1), 113–125 (1994) 4
14. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention U-Net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018) 2
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015) 3

16. Silva, B.M., Fonseca, J.M., Pinheiro, P., et al.: Artificial intelligence in dentistry: A comprehensive review. Journal of Dental Research **99**(9), 1010–1018 (2020) 1, 2
17. Van Nistelrooij, N., Bolelli, F., Vinayahalingam, S., Grana, C., et al.: ToothSeg: Robust tooth instance segmentation and numbering in CBCT using deep learning and self-correction. IEEE Journal of Biomedical and Health Informatics (2025), early access 2, 8
18. Wang, Y., Li, Z., Wu, C., Liu, J., Zhang, Y., Ni, J., Luo, Q., Chen, J., Zhang, H., Liu, J., et al.: Miccai sts 2024 challenge: Semi-supervised instance-level tooth segmentation in panoramic x-ray and cbct images. Medical Image Analysis p. 103986 (2026) 5, 9
19. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Dan, R., Zhang, Q., et al.: Miccai 2023 sts challenge: A retrospective study of semi-supervised approaches for teeth segmentation. Pattern Recognition **170**, 112049 (2026) 5, 9
20. Xu, Z., Escalera, S., Pavão, 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), 100543 (2022) 9
21. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage **31**(3), 1116–1128 (2006) 7
22. Zhang, Y., Wang, L., Chen, X., et al.: Multi-task learning for automated tooth instance segmentation and numbering in CBCT images. Medical Physics **52**(1), 455–468 (2025) 8
23. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning skip connections to exploit multiscale features in medical image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019) 2