Efficient CBCT Segmentation via nnU-Net with Structure-Aware Post-processing and Interactive Refinement

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

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University,
Shanghai 200240, People's Republic of China
{changkaiji, lswang}@sjtu.edu.cn

Abstract. Accurate segmentation of anatomical structures from cone-beam computed tomography (CBCT) is essential for clinical applications in dentistry, maxillofacial surgery, and orthodontics. The ToothFairy3 Challenge has a comprehensive 77-class segmentation task, emphasizing both accuracy and computational efficiency. In this work, we present a method based on the nnU-Net framework, enhanced with a Structure Aware Post-processing (SAP) strategy. nnU-Net serves as a backbone for multi-class segmentation, while SAP refines predictions by introducing individualized thresholds for each anatomical structure, thereby mitigating noise and preserving clinically important fine structures. To further improve efficiency, we disabled mirroring augmentation during training and employed inference acceleration strategies, including the removal of test-time augmentation and optimized interpolation on floating-point tensors. Experimental results validate the effectiveness of our approach in balancing segmentation accuracy with computational efficiency. To further ensure robustness in challenging clinical scenarios, we also utilize an interactive refinement module based on nnInteractive. This strategy allows clinicians to correct local segmentation errors with minimal user guidance, providing a safety net for complex anatomical variations.

Keywords: nnU-Net · Structure Aware Post-processing · Computational efficiency

1 Introduction

Cone-beam computed tomography (CBCT) has become an indispensable imaging modality in dentistry, maxillofacial surgery, and orthodontics due to its short acquisition time, low radiation dose, and high spatial resolution for hard tissues [13,1]. Accurate delineation of anatomical structures from CBCT is essential for surgical planning, risk assessment, and clinical decision-making. Building on the success of previous ToothFairy challenges, the ToothFairy3 – MICCAI 2025 competition pushes the boundaries of multi-class segmentation with an expanded dataset encompassing 77 anatomical categories, including newly introduced structures such as the pulp cavity, incisive nerve, and lingual foramen. This task emphasizes not only segmentation accuracy but also computational efficiency, reflecting the growing demand for real-time, reliable clinical tools.

Traditionally, the identification and delineation of anatomical structures in CBCT images have relied heavily on manual segmentation by experienced radiologists and dental professionals. The process requires substantial expertise and can take considerable time per case, making it impractical for routine clinical workflows where rapid decision-making is essential [16,5]. Moreover, the subjective nature of manual segmentation can lead to inconsistent results across different practitioners, potentially affecting treatment planning reliability.

In recent years, artificial intelligence techniques, particularly deep learning-based approaches using convolutional neural networks (CNNs), have demonstrated remarkable success in medical image segmentation tasks [15,10,9,17]. These automated methods have shown promising results in various dental imaging applications, offering the potential to significantly reduce processing time while maintaining or even improving segmentation accuracy. Deep learning frameworks have proven particularly effective at learning complex patterns and features from medical images, enabling robust identification of anatomical structures across diverse patient populations and imaging conditions [2,11,8].

Despite these advances, significant challenges remain for the ToothFairy3 task. First, the large number of categories (77) introduces class imbalance, as certain anatomical structures are underrepresented compared to larger, more prominent ones such as the mandible. This imbalance risks biasing the model toward dominant classes. Second, fine-scale structures like the incisive nerve or lingual foramen are difficult to segment reliably, requiring high-resolution features without overwhelming memory usage. Third, efficiency must be considered alongside accuracy: prolonged inference times or excessive memory consumption may render otherwise accurate models impractical for real-world clinical use [6,7]. Striking a balance between precision and computational efficiency is therefore essential.

To address these challenges, we propose a solution based on nnU-Net, enhanced with Structure Aware Post-processing (SAP) [14,4]. nnU-Net provides a strong backbone for multi-class segmentation, automatically adapting to the CBCT dataset's characteristics, while SAP refines predictions by removing spurious regions and ensuring anatomical plausibility. This approach aims to achieve high segmentation accuracy across 77 classes while maintaining computational efficiency, aligning with the dual objectives of the ToothFairy3 challenge. Despite the high performance of automated models, purely automatic segmentation may still falter in cases with severe artifacts or ambiguous boundaries (e.g., discontinuous inferior alveolar canals). To address this, we incorporate an interactive segmentation paradigm as a complementary refinement step. By leveraging user-provided point prompts, this module enables precise correction of difficult targets, ensuring that the system meets the rigorous reliability standards required for surgical planning. The contributions of our work can be summarized as follows:

- We employed an automated segmentation framework based on nnU-Net with SAP to address the multi-class segmentation challenge in CBCT images.
- The proposed approach optimizes the trade-off between segmentation quality and computational efficiency, ensuring both clinical accuracy and practical feasibility.
- Our approach achieved top-three performance in the ToothFairy3 Challenge validation phase, demonstrating its effectiveness for comprehensive dental and maxillofacial structure segmentation.

2 Proposed Method

2.1 Framework Overview

As shown in Fig. 1, we propose a segmentation approach for CBCT images, leveraging nnU-Net as the foundational architecture with disabled mirroring augmentation, combined with a SAP strategy. Disabling mirroring augmentation preserves the inherent left-right anatomical asymmetry of oral structures, enabling the model to learn structure-specific positional features. Structure Aware Thresholds provide adaptive morphological optimization, thereby minimizing false positives across diverse oral tissues.

2.2 Data Preprocessing

We employed nnU-Net's automated preprocessing pipeline to optimize data handling and network configuration for our multi-structure segmentation task. The preprocessing stage involved comprehensive dataset validation to ensure annotation consistency and data integrity across all CBCT volumes. The framework automatically determined optimal patch sizes, spacing parameters, and intensity normalization strategies based on the inherent characteristics of the dataset. This automated approach eliminates manual hyperparameter tuning while ensuring that preprocessing parameters are specifically tailored to the morphological and intensity characteristics of CBCT imaging data.

Additionally, the preprocessing pipeline established network topology and memory allocation strategies optimized for 3D volumetric segmentation of high-resolution CBCT images. The intensity normalization was performed using dataset-specific statistics computed from foreground regions, ensuring consistent intensity distributions across the training cohort.

2.3 Model Training Strategy

Model training was conducted using the 3D full-resolution configuration to preserve high spatial resolution critical for accurate delineation of fine anatomical

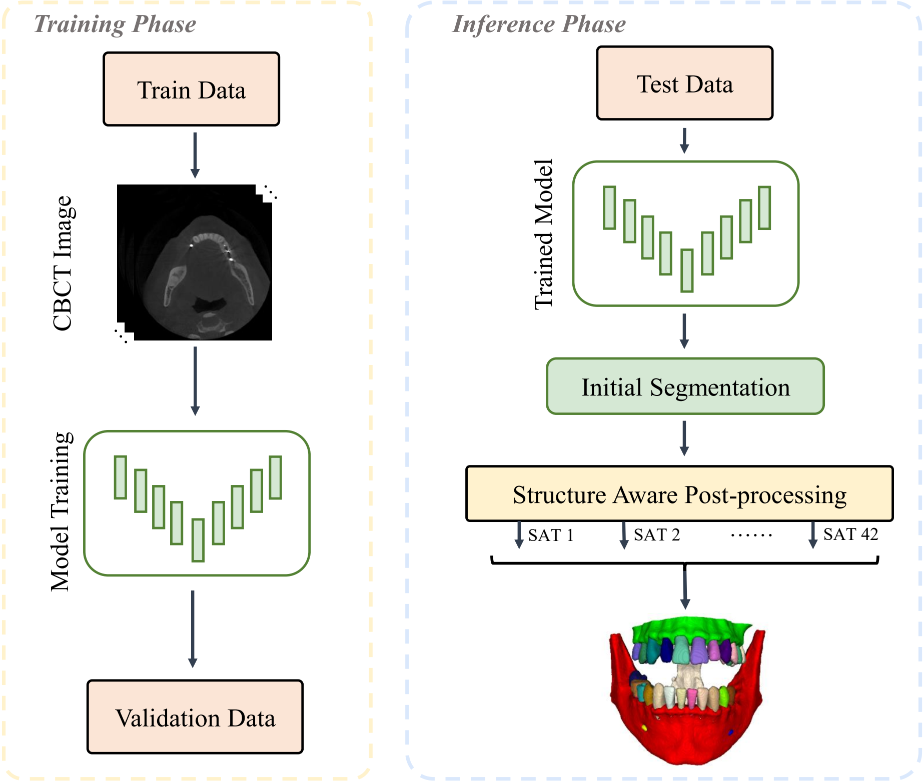

Fig. 1. Overview of the proposed framework. The training procedure utilized nnU-Net with disabled mirroring augmentation to enhance structure-specific learning. During inference, initial segmentation outputs were refined through Structure Aware Post-processing, wherein predetermined Structure Aware Thresholds were applied to target anatomical structures for morphological optimization.

structures. During the training phase, we selectively disabled mirroring-based data augmentation techniques. This approach addresses the inherent positional specificity of oral anatomical structures, where spatial location serves as one of the fundamental identifying characteristics. For instance, the left and right inferior alveolar canals, while morphologically similar, are distinguished primarily by their anatomical position. Similarly, FDI numbering assignment for teeth would be compromised by mirroring augmentation. Applying mirroring augmentation would artificially transpose these position-dependent structures, compromising the model's ability to learn spatial-anatomical relationships essential for accurate structure identification and increasing classification difficulty between bilaterally symmetric yet distinct anatomical entities.

2.4 Structure Aware Post-processing

Traditional post-processing approaches for medical image segmentation typically employ fixed filtering parameters across all anatomical structures, such as removing connected components smaller than a predetermined volume threshold or retaining only the largest connected component for each structure. However, this "one-size-fits-all" strategy presents significant limitations: overly conservative thresholds may preserve noise and erroneous segmentations, while aggressive thresholds risk eliminating clinically important small structures.

To address these limitations, we propose a Structure Aware Post-processing method that computes individualized Structure-aware Thresholds (SAT) for each anatomical structure, rather than applying uniform criteria across all structures. The core insight is that different anatomical structures exhibit distinct morphological characteristics and volume distributions, necessitating structure-specific optimization strategies.

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_K\}$ denote the set of K target anatomical structures. For each structure S_i, we define a SAT τ_i that specifies the minimum volume required for a connected component to be considered valid. The collection of thresholds is represented as $\mathbf{T} = \{\tau_1, \tau_2, \ldots, \tau_K\}$. Given an initial segmentation prediction $\mathbf{P}$, the structure-aware post-processing procedure consists of four sequential steps.

1. Connected Component Analysis. For each structure S_i, we extract all connected components from the corresponding segmentation mask:

$$\mathcal{C}_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,n_i}\},$$

where n_i denotes the number of connected components predicted for structure S_i, and each pair of components is disjoint, i.e., $c_{i,a} \cap c_{i,b} = \varnothing$ for $a \neq b$.

2. Volume Computation. For each connected component $c_{i,j}$, we compute its volume $v_{i,j}$. In the general case with voxel volume $V_{\text{vox}}(x, y, z)$, the volume is:

$$v_{i,j} = \sum_{(x,y,z) \in c_{i,j}} V_{\text{vox}}(x, y, z).$$

3. Threshold-based Filtering. Each connected component is retained only if its volume exceeds the corresponding threshold:

$$c_{i,j}^{\text{filtered}} = \begin{cases} c_{i,j}, & \text{if } v_{i,j} \geq \tau_i, \\ \varnothing, & \text{otherwise,} \end{cases}$$

where $\varnothing$ denotes the empty set, i.e., the component is discarded.

4. Final Reconstruction. The refined segmentation for structure S_i is obtained by the union of all retained components:

$$\hat{S}_i = \bigcup_{j:\, v_{i,j} \geq \tau_i} c_{i,j}.$$

The complete post-processed segmentation is given by

$$\hat{\mathbf{P}} = \{\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_K\},$$

which can be further represented as a labeled mask for downstream evaluation or visualization.

This approach enables differentiated treatment of anatomical structures with varying size characteristics. For instance, large structures such as jawbones can utilize higher thresholds to effectively eliminate substantial noise regions, while smaller structures like nerve canals employ lower thresholds to preserve their inherently compact morphology. The structure-aware post-processing thus provides a framework for balancing the trade-off between noise removal and structure preservation in multi-class anatomical segmentation tasks.

To determine the optimal values for the structure-aware thresholds (τ_{vol}), we analyze the volumetric distribution of each anatomical class within the training dataset. The filtering strategy is empirically tailored to the scale of the target structures. For the pharynx, we retain only the largest connected component. For other structures, thresholds are stratified by anatomical size: massive bone structures like the lower jawbone and upper jawbone utilize high thresholds ($10,000$ and $5,000$ voxels, respectively) to filter out major misclassifications. Medium-sized prosthetics employ a threshold of $2,000$ voxels. Specific subsets of teeth are assigned a threshold of $1,500$ voxels. Fine-grained structures, including the inferior alveolar canals, use a lower threshold of 500 voxels. The detailed configuration is presented in our Github.

2.5 Interactive Refinement Module

While the proposed nnU-Net with SAP achieves efficient automated segmentation, we introduce an interactive refinement module, nnInteractive, to handle corner cases requiring human expertise. This module adopts a "human-in-the-loop" workflow where clinicians can iteratively refine segmentation results using point prompts.

Network Architecture: Unlike methods using separate image and prompt encoders (e.g., SAM), nnInteractive employs an early prompt strategy. User-provided prompts (e.g., foreground/background clicks) are encoded as Gaussian heatmaps and concatenated with the original image and the current segmentation mask along the channel dimension. The network input consists of eight channels: the original image, the previous mask, and six channels representing different interaction types (points, scribbles, bounding boxes).

AutoZoom Mechanism: To handle small, fine-grained structures like the inferior alveolar canal within large FOV CBCT scans, the module incorporates an

AutoZoom mechanism. This dynamic strategy automatically crops and resamples the Region of Interest (ROI) around the user's interaction points, allowing the model to focus on local details at higher resolution without losing context. This ensures that even subtle anatomical structures can be precisely corrected with minimal user interaction (1–5 clicks).

3 Experiments and Results

3.1 Dataset and Assessment Metrics

The dataset used in Task 1 of the ToothFairy3 challenge is composed of CBCT scans annotated with 77 anatomical classes, encompassing not only large bony structures such as the mandible and maxilla, but also fine-grained elements such as pulp cavities, incisive canals, and the lingual foramen [3,2,12]. The volumes are provided in NIfTI format with intensity values in Hounsfield units. Across all scans, the maximum spatial dimensions are $(298, 512, 512)$, the minimum are $(170, 272, 345)$, and the median shape is $(168, 362, 371)$.

For evaluation, we adopt two widely used metrics in medical image segmentation: the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95). Both metrics are computed for each class on each test volume, followed by averaging across all volumes. DSC quantifies the voxel-wise overlap between the predicted segmentation and the ground truth, while HD95 assesses the boundary-level agreement by measuring the distance between surfaces. Together, these metrics capture both volumetric and geometric accuracy.

Although our analysis in this work focuses on DSC and HD95, it is worth noting that computational efficiency plays a crucial role in the challenge design. Inference runtime and maximum memory usage are also recorded and will contribute to the final ranking of submitted methods, reflecting their practical applicability in clinical settings.

3.2 Implementation details

Environments and Requirements. The training of our method was conducted for a total of 1000 epochs. The details of the computational environment and dependencies are summarized in Table 1.

Inference Acceleration. Since runtime was an important factor in the challenge evaluation, we applied several strategies to accelerate inference. First, we disabled test-time augmentation in nnU-Net, which substantially reduced the computational burden while maintaining competitive accuracy. Second, we optimized the handling of multi-class predictions by refining the interpolation step. Instead of relying on conventional integer-based resampling methods that are computationally demanding, we leveraged PyTorch's `interpolate` function on floating-point tensors. This choice preserves numerical precision while improving throughput in large-scale volumetric segmentation. Together, these strategies enabled efficient inference across the entire test set.

Table 1. System Configuration

Ubuntu version	Ubuntu 24.04 LTS
CPU	Intel(R) Xeon(R) Platinum 8352S CPU @ 2.20GHz
RAM	503 GB
GPU	1 NVIDIA GeForce RTX 4090 (24G)
CUDA version	12.4
Programming language	Python 3.9.19
Deep learning framework	PyTorch (torch 1.12.1, torchvision 0.19.1)
Code will available at	https://github.com/duola-wa/Toothfairy3

3.3 Results and Analysis

Quantitative Performance. The quantitative results for both the debug and test phases are summarized in Table 2. We report the Dice similarity coefficient and the HD95, with the former reflecting overlap accuracy and the latter assessing boundary alignment. Higher Dice and lower HD95 values indicate better performance.

Table 2. Evaluation results across debug and test phases. Dice similarity coefficient and HD95 are reported.

Metric	Statistic	Debug Phase	Test Phase
Dice Average	Min	0.9090	0.5671
	25%	0.9371	0.7340
	50%	0.9653	0.7821
	75%	0.9695	0.8329
	Max	0.9737	0.8670
	Mean	0.9493	0.7705
	Std	0.0352	0.0754
HD95 Average	Min	11.13	54.58
	25%	11.16	77.29
	50%	11.18	93.36
	75%	28.13	122.91
	Max	45.07	206.55
	Mean	22.46	104.59
	Std	19.58	37.21

In the debug phase, which included only three cases, our method demonstrated high segmentation accuracy with an average Dice score of 0.949 and a relatively low HD95 of 22.46. However, the larger-scale test phase presented more challenging scenarios, where the average Dice dropped to 0.770, and the mean HD95 increased to 104.59. This performance gap highlights the difficulty of generalization from a limited validation set to a more diverse and comprehensive test set. Nevertheless, the results remain competitive and validate the robustness of our approach under varying anatomical and imaging conditions.

Qualitative Results. To provide visual insight into the segmentation performance, Fig. 2 presents representative examples from the debug phase. These three cases illustrate the method's ability to accurately delineate anatomical structures across different imaging conditions and patient anatomies. Each row displays the input CBCT image (left), the predicted segmentation result (center), and the corresponding ground truth annotation (right). The visual comparison demonstrates the accuracy of our segmentation results on debug data, with predicted boundaries closely matching the expert annotations across multiple anatomical regions.

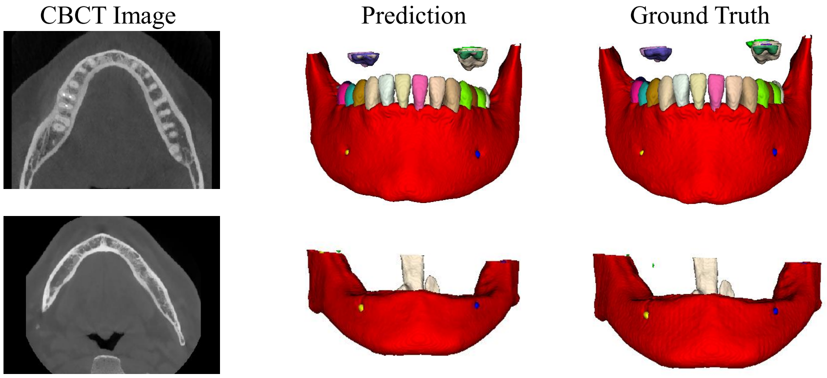

Fig. 2. Representative segmentation results from debug phase cases. Each row shows (from left to right): input CBCT image, predicted segmentation result, and ground truth. The results demonstrate accurate delineation of anatomical structures across different patient anatomies and imaging conditions on debug data.

As shown in Fig. 3, we provide a visual comparison of the segmentation results for nnInteractive with the introduction of 3 and 5 interaction points, respectively. This visualization highlights the impact of increasing the number of user interactions on the segmentation accuracy. Due to the limited number of submissions in the competition, we did not include metric-based results in this analysis, focusing instead on the visual comparison of the segmentation outputs.

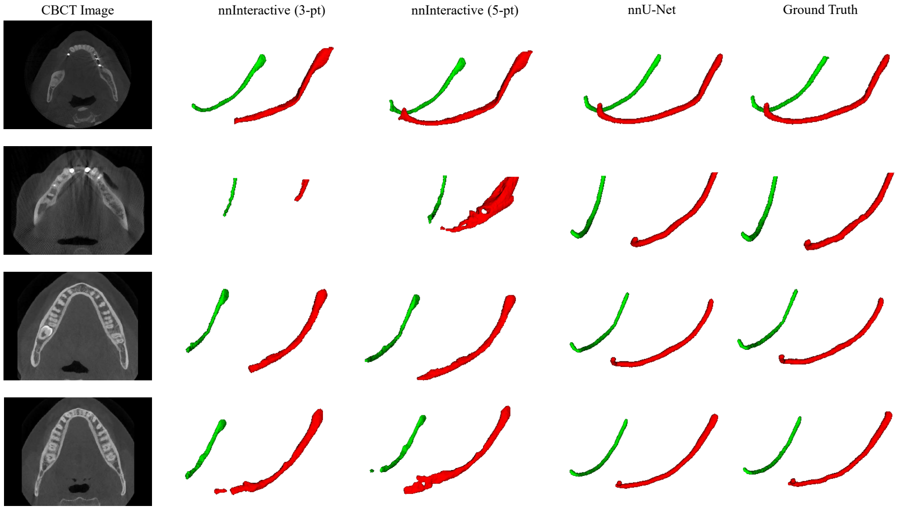

Fig. 3. The nnInteractive method is evaluated with 3 and 5 interaction points, highlighting the effect of prompt refinement on segmentation accuracy.

4 Conclusion

In this paper, we presented a segmentation framework for multi-class CBCT images, designed for the ToothFairy3 Challenge. Our approach leverages nnU-Net as a backbone and introduces SAP to account for the morphological variability of different anatomical structures. The proposed strategy enables differentiated handling of large and fine-scale structures, thereby reducing false positives while preserving clinically relevant details. Experiments demonstrated that our method achieves consistently high accuracy in the debug phase. Importantly, by optimizing interpolation strategies, we achieved notable improvements in inference efficiency. Furthermore, the integration of the interactive refinement module demonstrates a viable path for clinical deployment. It bridges the gap between fully automated processing and the need for meticulous precision in complex surgical cases, effectively balancing algorithmic efficiency with clinical reliability.

References

1. Acar, B., Kamburoğlu, K.: Use of cone beam computed tomography in periodontology. World journal of radiology **6**(5), 139 (2014)
2. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., Van Nistelrooij, N., Van Lierop, P., Xi, T., Liu, Y., et al.: Segmenting the inferior alveolar canal in cbcts volumes: the toothfairy challenge. IEEE Transactions on Medical Imaging (2024)
3. 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 the Computer Vision and Pattern Recognition Conference. pp. 5238–5248 (2025)

4. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)
5. Ji, C., Du, C., Zhang, Q., Wang, S., Ma, C., Xie, J., Zhou, Y., He, H., Shen, D.: Mammo-net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 68–78. Springer (2023)
6. Ji, C., Liu, Y., He, L., Jiang, Y., Huang, C., Wang, L.: Two-stage semi-supervised nnu-net framework for tooth segmentation in cbct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 100–109. Springer (2024)
7. Ji, C., Liu, Y., He, L., Jiang, Y., Huang, C., Wang, L.: A two-stage semi-supervised nnu-net model for automated tooth segmentation in panoramic x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 91–99. Springer (2024)
8. Jiang, Y., Liu, Y., Ji, C., Wang, L.: Enhanced multi-structure segmentation in cbct images with adaptive structure optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 30–40. Springer (2024)
9. Lin, Z., Liu, Y., Wu, J., Wang, D.H., Zhang, X.Y., Zhu, S.: Multi-modal pre-post treatment consistency learning for automatic segmentation and evaluation of the circle of willis. Computerized Medical Imaging and Graphics **122**, 102521 (2025)
10. Liu, Y., Xin, R., Yang, T., Wang, L.: Inferior alveolar nerve segmentation in cbct images using connectivity-based selective re-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2024)
11. Liu, Y., Zhao, Z., Wang, L.: A cnn-based multi-stage framework for renal multi-structure segmentation. In: MICCAI Challenge on Correction of Brainshift with Intra-Operative Ultrasound, pp. 18–26. Springer (2022)
12. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing patch-based learning for the segmentation of the mandibular canal. IEEE Access **12**, 79014–79024 (2024)
13. Patel, S., Durack, C., Abella, F., Shemesh, H., Roig, M., Lemberg, K.: Cone beam computed tomography in e ndodontics–a review. International endodontic journal **48**(1), 3–15 (2015)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
15. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**(1), 221–248 (2017)
16. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)
17. Yang, T., Yu, X., Tao, R., Li, H., Zhou, J.: Blood glucose prediction for type 2 diabetes using clustering-based domain adaptation. Biomedical Signal Processing and Control **105**, 107629 (2025)