# Efficient and Robust CBCT Segmentation of Oral and Maxillofacial Structures

Fan Xiao[1], Xinrui Huang[2], Anqi Gao[1,3], Dongming He[1], Xiaofan Zhang[2], and Xudong wang[1,3,4,5,6,7,8(✉)]

[1] Department of Oral Craniomaxillofacial, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
ff741333@gmail.com,xudongwang70@hotmail.com
[2] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
{huangxr,xiaofan.zhang}@sjtu.edu.cn
[3] College of Stomatology, Shanghai Jiao Tong University, Shanghai, China
[4] National Center for Stomatology, Shanghai, China
[5] National Clinical Medical Research Center for Oral Diseases, Shanghai, China
[6] Shanghai Key Laboratory of Stomatology, Shanghai, China
[7] Shanghai Research Institute of Stomatology, Shanghai, China
[8] Research Unit of Oral and Maxillofacial Regenerative Medicine, Chinese Academy of Medical Science, Shanghai, China

**Abstract.** In dental practice, accurate segmentation of oral and maxillofacial structures from cone-beam computed tomography (CBCT) images is essential for diagnostic and treatment planning purposes. However, manual segmentation is time-consuming and labor-intensive. Although numerous deep learning-based methods have been proposed to automate this process, most rely on a single model architecture, which struggles to handle the complex and diverse nature of oral anatomical structures. To address this limitation, we propose a hybrid framework integrating nnUNet and VISTA models for automated and interactive segmentation of oral and maxillofacial structures. Our approach employs a class-wise ensemble strategy to improve inference efficiency and accuracy, and incorporates post-processing techniques such as threshold-based small object removal and disconnected region filtering to enhance robustness. The proposed method achieved third place in Task 1 and second place in Task 2 of the ToothFairy3 Challenge. Code and model weights are available at https://github.com/ff741333/toothfairy3_blcakmyth.

**Keywords:** CBCT image · Oral and maxillofacial structures segmentation · Interactive segmentation

## 1 Introduction

In dental practice, obtaining accurate oral and maxillofacial structures is essential for disease diagnosis and treatment. Cone-beam computed tomography

(CBCT), as a commonly used imaging modality, is frequently applied in dentistry and related fields due to its advantages of short acquisition time, low radiation dose, and high resolution for hard tissues. The oral and maxillofacial structures that can be obtained from CBCT images are illustrated in Figure 1, including teeth (and dental attachments such as bridges, crowns, and implants), jawbone, maxillary sinus, pharynx, inferior alveolar canal (IAC), mandibular incisive canal, lingual canal, and others. These anatomical structures are critical for clinical applications such as surgical planning in implantology [6] and maxillofacial surgery [9], as well as tooth alignment in orthodontics. However, manually segmenting these structures from CBCT images is time-consuming and labor-intensive.

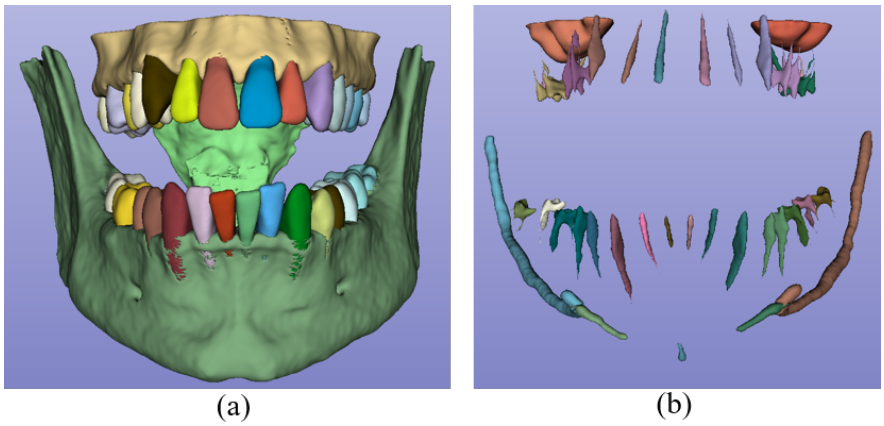


**Fig. 1.** Visualization of Oral and Maxillofacial Structures in CBCT: (a) Jawbone, Teeth, Pharynx; (b) Inferior Alveolar Canal, Mandibular Incisive Canal, Lingual Canal, Maxillary Sinus, Pulp

In recent years, numerous studies [4,16,5,2,8] have focused on achieving automatic segmentation of oral and maxillofacial structures. Cui et al. [4] proposed a two-stage deep network leveraging hierarchical tooth morphology for precise tooth segmentation and a filter-enhanced network enhancing intensity contrasts for accurate alveolar bone segmentation. Dot et al. [5] has developed an open-source tool for robust segmentation of oral and maxillofacial structures on CBCT and CT images, including the maxilla, mandible, teeth, and mandibular canal. Bolelli et al. [2] constructed a dataset consisting of 42 different types of CBCT maxillofacial structure segmentation, and employed various strategies to optimize the performance of existing excellent segmentation models [10,3,18,15,14,12,11]. The existing segmentation models for oral and maxillofacial structures are usually based on a single architecture, which is insufficient for their complex and diverse nature. Each tooth is of a similar size, yet they vary in morphology and position. Additionally, the jawbone has a relatively large

volume and contains numerous neural structures. Different model architectures possess varying receptive fields and exhibit differences in segmenting diverse oral and maxillofacial structures. Therefore, designing diverse model architectures is highly beneficial for the segmentation of oral and maxillofacial structures.

In this work, we propose an algorithm for segmenting different oral and maxillofacial structures in CBCT images based on the nnUNet [10] and VISTA [7] framework, which also supports interactive segmentation of the IAC. To balance inference efficiency and accuracy, we designed multiple strategies to optimize the algorithm's inference process. Additionally, we employed post-processing techniques such as custom threshold-based small label removal and non-connected region filtering to further enhance robustness. Finally, we validated our algorithm in the ToothFairy3 Challenge, achieving **3**$^{\text{rd}}$ place on Task 1 and **2**$^{\text{nd}}$ place on Task 2.

## 2 Method

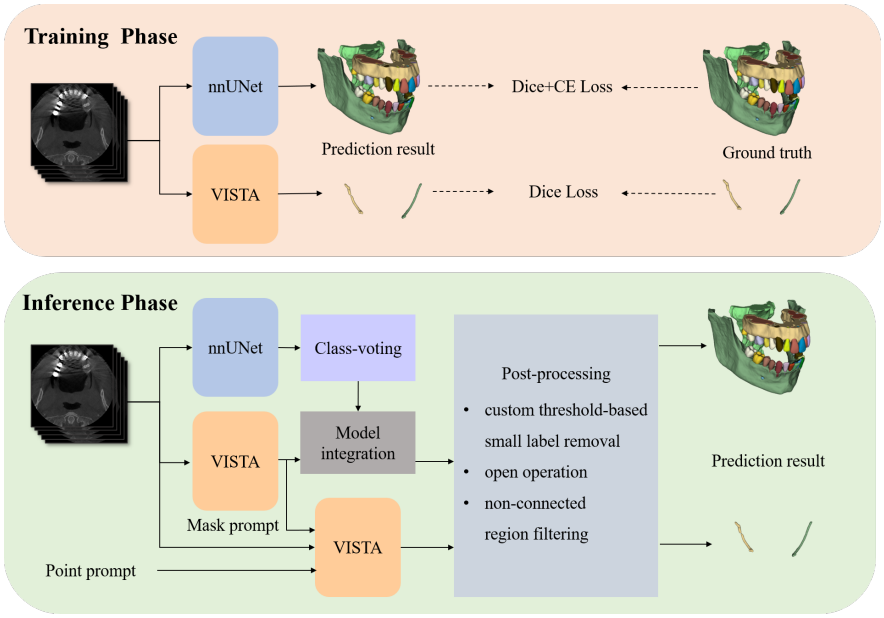


**Fig. 2.** The framework of our proposed method. An nnUNet model is trained for oral and maxillofacial segmentation and a VISTA model for interactive IAC segmentation. Inference combines the models with a class-voting ensemble and a two stage automatic and point-based refinement for interactive segmentation. Outputs of both models are post-processed to improve robustness.

Our proposed method is illustrated in Figure 2. We first separately train an nnUNet-based model for oral and maxillofacial structures segmentation and a VISTA-based model for interactive IAC segmentation. Considering computational efficiency, we adopt a class-voting ensemble strategy to reduce memory usage and improve inference speed. To take advantage of the differences in various oral and maxillofacial structures, we integrate the two models to enhance segmentation accuracy. For the interactive segmentation task, we adopt a two-stage strategy: an automatic segmentation stage is first applied to obtain an initial result, followed by a point-based prompt refinement stage to further enhance the segmentation accuracy. Finally, we apply post-processing techniques such as custom threshold-based small label removal and non-connected region filtering to the outputs of both models to further enhance robustness.

### 2.1 Model training

For nnUNet model training phase, we chose nnUNet ResEnc L as backbone network. Since all data share the same spacing [0.3, 0.3, 0.3], no resampling was performed on any data. The data were cropped into patches of size 128×224×224 and augmented using techniques including rotation within the range of $-30°$ to $30°$, scaling with a factor of 0.7–1.4, anterior–posterior and superior–inferior mirroring, addition of Gaussian noise (variance 0.1) and Gaussian blur (sigma 0.5–1.0), and contrast adjustment with a factor of 0.75–1.25. The training proceeded for 1500 epochs with a batch size of 2. The SGD optimizer was adopted with an initial learning rate of 0.01 and a Poly scheduler [17]. The loss function was defined as the sum of Dice Loss and Cross-Entropy Loss.

For VISTA model training phase, we chose VISTA3D pretrained checkpoint. No resampling was also performed on any data. The data were cropped into patches of size 96×160×160 and augmented using techniques including scaling with a factor of 0.8–1.2, simulation of low resolution images with a factor of 0.3–1.0, addition of Gaussian noise (variance 0.2) and Gaussian blur (sigma 0.5–1.0), and contrast adjustment with a factor of 0.9–1.1. The training proceeded for 200 epochs with a batch size of 2. The AdamW optimizer was adopted with an initial learning rate of 5e-5, a weight decay of 1e-5 and a cosine scheduler. The loss function was defined as Dice Loss.

We utilized all the data provided in Toothfairy3 [1,2,13] as both the training set and the test set, without performing any data partitioning. However, while the nnUNet model employed all available labels, the VISTA model only used the labels corresponding to the IAC.

### 2.2 Ensemble strategy

**Class-voting** The sliding window inference process of the nnU-Net model is highly memory-intensive. To achieve efficient segmentation, we adopt a class-voting strategy. Specifically, for each patch $S_i$, the predicted logits are converted into one-hot encoded vectors, which are then aggregated through summation.

The final predicted label is determined by selecting the class with the maximum accumulated value. The formulation is as follows:

$$\hat{y} = \arg \max_{c \in 1,\cdots,C} \sum_{i=1}^{N} \mathbb{I}_c \left( \arg \max f(S_i) \right) \tag{1}$$

where $\hat{y}$ denotes the final predicted label, $C$ represents the total number of classes, $N$ is the total number of sliding window patches that cover the spatial location, $f(S_i)$ denotes the predicted logits for patch $S_i$, $\mathbb{I}_c(\cdot)$ is the indicator function that outputs 1 if the argument equals class $c$ and 0 otherwise, $\arg \max f(S_i)$ obtains the predicted class label for patch $S_i$. The outer $\arg \max$ operation selects the class with the highest vote count.

**Model integration** The VISTA model demonstrated superior performance compared to the nnUNet model in segmenting the IAC. Therefore, we integrated the nnUNet model with the VISTA model. Specifically, we first removed the IAC segmentation labels predicted by the nnUNet model and then replaced them with the corresponding labels generated by the VISTA model. This approach allowed us to substitute the relatively inaccurate IAC labels from nnUNet with more precise ones.

### 2.3 Interactive segmentation

To achieve better interactive segmentation results, we employ a method that combines automatic segmentation with point-prompt-based interactive segmentation. Specifically, we first use the automatic segmentation decoder from the VISTA model to obtain the automatic segmentation mask of the IAC, and then apply the point-prompt decoder to generate the point-prompt-based segmentation mask. We then add or remove only the connected component regions that contain the point clicks to avoid unexpected modifications. This refinement of the automatic segmentation results using point-prompt-based segmentation significantly improves the overall performance.

### 2.4 Post-processing techniques

The post-processing techniques employed in our pipeline are designed to enhance segmentation accuracy and robustness by incorporating both morphological operations and prior anatomical knowledge. These techniques include custom threshold-based small label removal, morphological open operations, and non-connected region filtering. Specifically, we first perform label size filtering to remove anatomically implausible labels whose pixel area falls within predefined ranges (e.g., 320–1,819 pixels for certain upper teeth and 970–6,140 pixels for wisdom teeth), as determined from the training set distribution. This step helps to suppress spurious predictions and reduce false positives associated with small, isolated regions.

Next, a morphological open operation is applied to the upper jawbone and pharynx regions to address potential boundary ambiguities and to smooth jagged edges in the predicted segmentation masks. The open operation, which consists of an erosion followed by a dilation, effectively removes small noise while preserving the overall structure of the anatomical regions.

Finally, non-connected region filtering is performed on the lower jawbone, pharynx, and tooth labels. This step leverages prior anatomical knowledge by retaining only the largest connected components for each label, thereby eliminating isolated or incorrectly segmented regions that are inconsistent with realistic anatomical structures. By combining these post-processing steps, our framework not only improves the visual consistency of the segmentations but also enhances quantitative metrics by reducing both false positives and false negatives in critical regions.

## 3 Experiment

### 3.1 Implementation details

The training of the nnUNet model was conducted on two NVIDIA GeForce RTX 4090 GPUs, while all other training and experiments were performed on a single NVIDIA GeForce RTX 4090 GPU. The metrics used in the experiment include Dice, HD95, and inference time.

| Models | Patch size | Dice | Prediction time(s) |
|---|---|---|---|
| nnWnet L | 96x160x160 | 0.7886 | 27.7 |
| nnWnet M | 96x160x160 | 0.7868 | 21.3 |
| nnWnet S | 96x160x160 | 0.7774 | 13.7 |
| nnUNet ResEnc L | 96x160x160 | 0.8037 | 5.0 |
| U-mamba | 128x224x224 | **0.8535** | 7.7 |
| nnUNet ResEnc L | 128x224x224 | 0.8312 | **4.7** |

**Table 1.** Comparison with other models. Prediction time refers to the time taken by the model to predict a single CBCT.

**Comparison with other models** In consideration of the balance between efficiency and accuracy of the algorithm, we compared several existing models, such as U-Mamba [14] and nnWNet [19]. As shown in Table 1, the Dice score of nnUNet ResEnc L (96×160×160) reaches 0.8037, which is higher than that of all nnWNet (ranging from 0.7774 to 0.7886), but lower than nnUNet ResEnc L (128×224×224) with Dice score of 0.8312. This indicates that enlarging the patch size contributes to performance improvement, as nnUNet ResEnc L with patches (128×224×224) outperform their counterparts trained on smaller patches (96×160×160). In terms of inference speed, nnUNet ResEnc L demonstrates the best efficiency, requiring only 4.7 seconds per CBCT, which is faster

than all other models including U-Mamba (7.7 seconds). These results suggest that nnUNet ResEnc L achieves a favorable trade-off, delivering the fastest inference while maintaining competitive segmentation accuracy.

| Class-voting | Model integration | Post-processing | Dice | HD95 | Inference time(s) |
|---|---|---|---|---|---|
| × | ✓ | ✓ | **0.9575** | **18.66** | 51.0 |
| ✓ | × | ✓ | 0.9563 | 18.67 | 25.5 |
| ✓ | ✓ | × | 0.8772 | 55.88 | **25.2** |
| ✓ | ✓ | ✓ | 0.9573 | **18.66** | 35.0 |

**Table 2.** Ablation study on debugging phase. Inference time denotes the complete duration required for processing a single CBCT during inference.

**Ablation study** To further investigate the contributions of different components, we conducted an ablation study, as summarized in Table 2. When only model integration and post-processing were employed, the framework achieved the highest Dice score of 0.9575, with an HD95 of 18.66, albeit at the cost of the longest inference time (51.0 s). By contrast, applying class-voting with post-processing but without model integration reduced the Dice score slightly to 0.9563 while improving efficiency (25.5 s). Removing post-processing led to a substantial degradation in accuracy, with the Dice dropping to 0.8772 and HD95 increasing to 55.88, although this configuration achieved the fastest inference (25.2 s). Incorporating all three components (class-voting, model integration, and post-processing) yielded a balanced performance, with a Dice of 0.9573, HD95 of 18.66, and moderate inference time (35.0 s). These results highlight the critical role of post-processing for maintaining segmentation accuracy and demonstrate that combining ensemble strategies can effectively balance accuracy and efficiency.

| Task | Team | Dice | HD95 |
|---|---|---|---|
| Multi-class Segmentation | Black_Myth | **0.7981±0.0640** | **88.7228±32.3250** |
| | TAIR Lab | 0.7917±0.0652 | 93.1873±30.4327 |
| | sjtu_eiee_2-426lab | 0.7705±0.0754 | 104.5936±37.2139 |
| | ring821 | 0.7684±0.0969 | 104.4004±47.9841 |
| | DLaBella29 | 0.7386±0.0708 | 97.7059±33.2051 |
| IAC Interactive Segmentation | Black_Myth | **0.8642±0.0507** | **2.2675±1.7112** |
| | TAIR Lab | 0.8519±0.0752 | 7.3863±20.4354 |
| | DLaBella29 | 0.7465±0.0724 | 4.7094±3.4713 |
| | sjtu_eiee_2-426lab | 0.7683±0.1896 | 32.2318±85.8957 |
| | gagaha | 0.7220±0.2554 | 76.8298±159.1321 |

**Table 3.** Final result on test phase leaderboards.

Finally, our final results are presented in Table 3. We achieved the best performance on both the Multi-class Segmentation leaderboard and the IAC Interactive Segmentation test phase leaderboard of the MICCAI Toothfairy3 Challenge. However, due to considerations regarding algorithmic runtime and computational cost, our final official standings were third place on Task 1 and second place on Task 2.

## 4 Discussion

This work presents a segmentation framework that integrates nnUNet and VISTA for accurate delineation of oral and maxillofacial structures in CBCT images. The complementary strengths of the two architectures—nnUNet for large-volume structures and VISTA for the fine-grained IAC—enabled superior performance compared with single-model approaches. The use of class-voting, interactive segmentation, and post-processing further enhanced efficiency and robustness, which was reflected in our third place result in Task 1 and second place result in Task 2 of the ToothFairy3 Challenge.

Nevertheless, the method was trained and validated only on the challenge dataset, and its generalizability to multi-center or clinical data remains to be verified. Future work will explore multi-institutional validation, lightweight deployment strategies, and extension to pathological segmentation for broader clinical applicability.

**Acknowledgments.** This study was funded by National Key R&D Program of China (2023YFC2414100), National Natural Science Foundation of China (82370905, 82071096), Shanghai Professional Service Platform of Oral-Cranio-Maxillofacial Digital Technology Research and Application (21DZ2294600), National Clinical Key Specialty (Z155080000004), Shanghai's Top Priority Research Center (2022ZZ01017), and CAMS Innovation Fund for Medical Sciences (CIFMS, 2019-I2M-5-037).

## References

1. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., R. Rokuss, M., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles Ballester, Vicent amd Paolo Burgos-Artizzu, X., Prados Carrasco, F., Berge', S., van Ginneken, B., Anesi, A., Grana, C.: Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. IEEE Transactions on Medical Imaging pp. 1–17 (Dec 2024). `https://doi.org/https://doi.org/10.1109/TMI.2024.3523096`
2. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volume. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–10. IEEE (Mar 2025)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

4. 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)
5. Dot, G., Chaurasia, A., Dubois, G., Savoldelli, C., Haghighat, S., Azimian, S., Taramsari, A.R., Sivaramakrishnan, G., Issa, J., Dubey, A., et al.: Dentalsegmentator: robust open source deep learning-based ct and cbct image segmentation. Journal of dentistry **147**, 105130 (2024)
6. Elgarba, B.M., Van Aelst, S., Swaity, A., Morgan, N., Shujaat, S., Jacobs, R.: Deep learning-based segmentation of dental implants on cone-beam computed tomography images: A validation study. Journal of Dentistry **137**, 104639 (2023)
7. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., et al.: Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography. CoRR (2024)
8. Huang, X., He, D., Li, Z., Zhang, X., Wang, X.: Iossam: Label efficient multi-view prompt-driven tooth segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 632–642. Springer (2024)
9. Huang, X., He, D., Li, Z., Zhang, X., Wang, X.: Maxillofacial bone movements-aware dual graph convolution approach for postoperative facial appearance prediction. Medical Image Analysis **99**, 103350 (2025)
10. 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)
11. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention. pp. 615–625. Springer (2024)
12. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)
13. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access pp. 1–12 (2024). https://doi.org/https://doi.org/10.1109/ACCESS.2024.3408629
14. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
15. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging **43**(9), 3377–3390 (2024)
16. Wang, Y., Xia, W., Yan, Z., Zhao, L., Bian, X., Liu, C., Qi, Z., Zhang, S., Tang, Z.: Root canal treatment planning by automatic tooth and root canal segmentation in dental cbct with deep multi-task feature learning. Medical image analysis **85**, 102750 (2023)
17. Zhang, T., Li, W.: kdecay: Just adding k-decay items on learning-rate schedule to improve neural networks. arXiv preprint arXiv:2004.05909 (2020)
18. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE transactions on image processing **32**, 4036–4045 (2023)
19. Zhou, Y., Li, L., Lu, L., Xu, M.: nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20852–20862 (June 2025)