A Multi-stage Network with Self-attention for Tooth Instance Segmentation

Yongcun Zhang, Zhiming Luo(✉), and Shaozi Li

Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China
{zhiming.luo,szlig}@xmu.edu.cn

Abstract. Automatic and accurate instance segmentation of teeth from 3D Cone-Beam Computer Tomography (CBCT) images is crucial for dental diagnose. Although Convolutional Neural Networks (CNNs) are widely used for tooth instance segmentation, the limitations of CNNs in capturing global image information can impact model performance. Recently, Transformer models leveraging the Self-Attention mechanism have exhibited exceptional capabilities in modeling global relationships in images. In this paper, we propose a fully automated tooth instance segmentation model utilizing the Self-Attention mechanism. The model is primarily based on the Self-Attention UNETR++ network and consists of three stages. In the first stage, a V-Net is employed to identify the region of interest (ROI) containing the teeth. In the second stage, a multitask UNETR++ network is utilized to extract the centroid and skeleton of the teeth. In the third stage, another multitask UNETR++ is employed to simultaneously learn the tooth mask and boundary, leading to accurate tooth instance segmentation. Experimental results on a dataset consisting of 98 CBCT images demonstrate the efficacy of our method. It achieves a Dice score of 95.1% and reduces the average surface distance (ASD) to 0.14 mm.

Keywords: Tooth segmentation · Self-Attention · CBCT image

1 Introduction

Currently, there is an increasing demand for dental health. Dental health issues mainly include dental diseases, dental implants, orthodontics. Although the growing number of patients seeking dental diagnoses contributed to the rapid development of the dental healthcare market, there is a significant shortage of dentists per million population, which poses a substantial burden on dentists. In clinical diagnosis, Cone Beam Computer Tomography (CBCT) is widely utilized for acquiring high-resolution 3D images of teeth, thereby offering accurate representations of dental crowns, roots, and bones. Additionally, CBCT offers the advantages of low Radiation exposure and short scanning time. On the other hand, the voxel information in CBCT images is highly complex, necessitating extensive manual segmentation to extract vital information. Therefore, this process becomes time-consuming and labor-intensive for clinicians and researchers.

Y. Sun et al. (Eds.): ChineseCSCW 2023, CCIS 2013, pp. 438–450, 2024.
https://doi.org/10.1007/978-981-99-9640-7_32

Thus, the development of digital dentistry and fully automated tooth segmentation methods is crucial for tooth analysis from 3D CBCT scans.

Computer vision technology has found widespread applications in the field of medical imaging. Driven by computer vision technology, digital oral cavity is rapidly developing. Automatic tooth segmentation is a primary step for tooth image analysis, and has attracted more and more research attention. Existing tooth instance segmentation methods can be categorized into two types: traditional methods and deep learning-based methods. Traditional methods, such as level set [1,14,15,20], graph cut [18,21], and template fitting [2,29]. However, these method rely on manually designed features, which are highly sensitive to complex dental situations, requiring tedious manual initialization and correction. They often lead to suboptimal segmentation performance in complicated cases. Deep learning methods, on the other hand, are known for their automatic feature extraction, strong adaptability, and high accuracy. They have been widely adopted in medical image segmentation.

Deep learning-based tooth instance segmentation methods [7–12,19,22,31] generally achieve better performance than traditional methods. However, nearly all deep learning-based methods rely on convolutional neural networks (CNNs) to extract features from CBCT images and achieve tooth detection and segmentation. None of these methods introduce attention mechanisms. CNNs' limitations in obtaining global image information to some extent lower the model's performance. Overcoming these limitations and improving the performance of tooth segmentation models pose challenging tasks. The widespread application of Transformers [5,13,25,33] indicates that Self-Attention can effectively obtain global information of images. This makes the model based on the Self-Attention mechanism have certain advantages in the field of image segmentation. In recent years, several outstanding neural networks using Self-Attention mechanisms [4,6,16,32,34] have emerged in the medical image segmentation field. Therefore, this paper aims to construct a fully automatic tooth instance segmentation method incorporating Self-Attention mechanisms.

Inspired by the above work, we propose a fully automated tooth instance segmentation model that utilizes the Self-Attention mechanism. The model has three stages: First, we use V-Net [24] to extract tooth ROI. Next, we use a multi-task UNETR++ network [28] to predict the centroids and skeletons of teeth. This step localizes teeth, detects tooth shapes, and represents teeth. Finally, we further segment teeth within the tooth ROIs using the multi-task UNETR++ network. By combining the centroids and skeletons of teeth, we achieve tooth instance segmentation. To evaluate the performance of our method, our fully automatic tooth segmentation achieved a Dice similarity coefficient of 95.1% and an Average Surface Distance of 0.14mm in tooth segmentation.

In summary, the main contribution of this study are as follows:

1. We propose a multi-stage model that is capable of fully automatic tooth instance segmentation on input 3D CBCT images.

2. By introducing a self attention mechanism, we have effectively improved the segmentation accuracy of our model which surpasses the performance of other comparative models.
3. By using multitasking learning, we successfully reduced the error in tooth surface segmentation while maintaining a high level of mask segmentation accuracy.
4. By evaluating our model with other CNN based models through experiments, we have demonstrated that introducing self attention mechanism can improve the performance of tooth segmentation models.

2 Related Work

Tooth Segmentation Based on Deep Learning. Inspired by 3D Mask R-CNN [17], Cui et al. [11] introduced ToothNet, an automatic tooth instance segmentation method in CBCT images. ToothNet employs 3D Region Proposal Network (RPN) [26] for tooth detection, recognition, and segmentation. Chung et al. [8] proposed the PATRCNN+TSNet method, which addressing metal artifacts in CBCT images using pose-aware techniques. Chen et al. [7] presented 3D FCN+MWT, a method that combines deep learning and traditional methods. They utilized a multi-task 3D fully convolutional network (FCN) to simultaneously predict tooth masks and surfaces. They then employed marker-controlled watershed transform (MWT) for tooth recognition and segmentation. Wu et al. [31] incorporated a center-sensitive mechanism into their method to guide tooth localization, thus avoiding the computational burden of numerous anchors generated by RPN in 3D CBCT images. Additionally, they employed DenseASPP-UNet for tooth segmentation and added boundary loss to reduce prediction errors on tooth boundaries. Jang et al. [19] proposed PanoramicNet, a novel tooth instance segmentation method. This method first expands the 3D tooth image into a 2D Panorama by calculating the dental arch curve. Then, it detects the teeth on the 2D Panorama images and completes instance segmentation by combining the 2D and 3D results. To address the diverse and complex tooth morphologies and reduce computational complexity, Cui et al. [12] extended their previous work [11] and introduced Hierarchical Morphology-Guided Network (HMGNet). The HMGNet utilizes tooth centroids to represent tooth positions and introduced tooth skeletons to depict the tooth's morphological structure, which can significantly enhance tooth segmentation accuracy in complex cases.

Self-attention. The Self-Attention mechanism calculates the similarities between different positions in the input sequence, assigns weights to each position, and then uses these weights to compute the output for each position. Specifically, given an input sequence X, it first performs linear transformations to obtain three matrices Q, K, V. Next, it calculates the similarity matrix QK^T by taking the dot product of each row vector in matrix Q and matrix K. Finally,

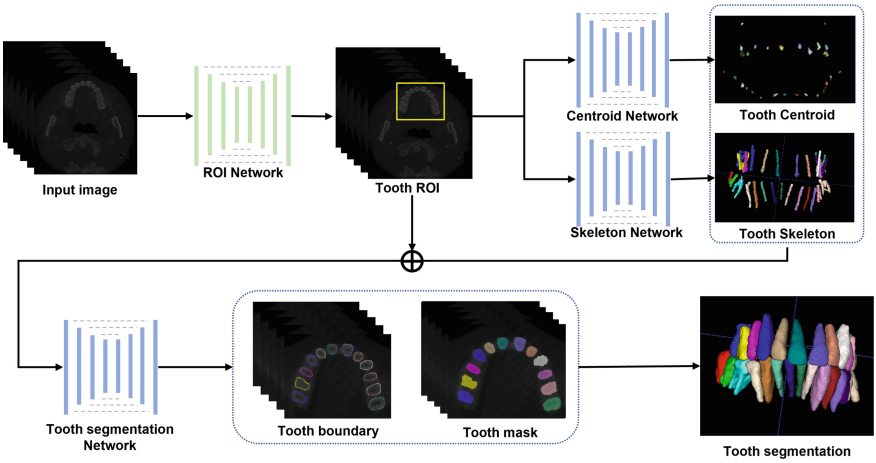

Fig. 1. The overall architecture of our method for fully automatic tooth instance segmentation.

these similarities are normalized into a probability distribution using the softmax function. The result is multiplied with matrix V to obtain the Self-Attention representation [30]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V, \tag{1}$$

where d_k represents the dimension of the key vector for stabilizing the learning process. Self-Attention allows the model to capture long-range dependencies and global information from the input sequence, which can lead to improved performance in various tasks. In this paper, we introduce the Self-Attention mechanism in tooth instance segmentation for improving the accuracy.

3 Method

The overall architecture of our method for fully automatic tooth instance segmentation is shown in Fig. 1, which mainly consists of three stages. In the first stage, V-Net [24] is employed for coarse binary segmentation of teeth to obtain the teeth Region of Interest (ROI). In the second stage, a muti-task UNETR++ is used to extract teeth centroids and skeletons. These provide a rough representation of the morphological structure of teeth. The third stage involves utilizing an another muti-task UNETR++ for tooth segmentation with the guidance of the tooth skeleton. This stage simultaneously generates teeth masks and boundaries.

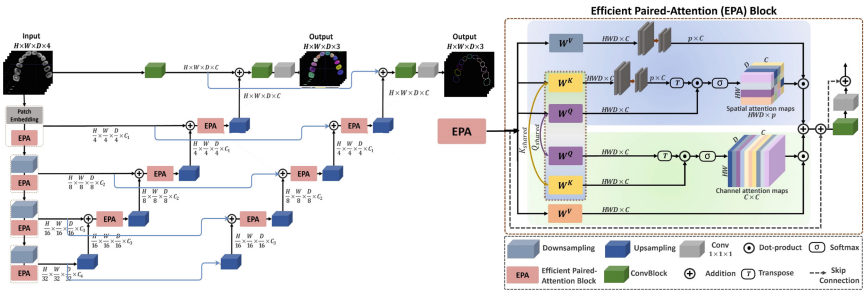

Fig. 2. The architecture of Multi-task UNETR++.

3.1 Multi-task UNETR++

In this paper, the multi-task UNETR++ is employed as the backbone network for tooth instance segmentation. As depicted in Fig. 2, the network follows a hierarchical Encoder-Decoder structure. To process the input 3D image, it is first converted into 3D patches using Patch Embedding [13]. Given an input image $X \in R^{H \times W \times D}$, it is partitioned into patches of resolution (P_h, P_w, P_d), resulting in feature maps of size $\frac{H}{P_h} \times \frac{W}{P_w} \times \frac{D}{P_d} \times C$. Throughout the experiments, a patch size of (4, 4, 4) is utilized. The designed multi-task UNETR++ introduces an additional decoder for the achieving the multi-tasks, such as tooth mask segmentation and tooth boundary estimation. Both decoders use skip connections to obtain feature maps from the encoder at each layer.

The core design of UNETR++ is the Efficient Pairwise Attention (EPA) blocks. It can effectively learn spatial and channel features through a pair of interdependent branches based on spatial and channel attention [28]. According to Eq. 1, spatial and channel attention can be calculated:

$$\begin{aligned} A_s &= Attention(Q_{shared}, K_{spatial}, V_{spatial}) \\ A_c &= Attention(Q_{shared}, K_{shared}, V_{channel}) \end{aligned} \tag{2}$$

In the spatial attention, $V_{spatial}(HWD \times C)$ and $K_{shared}(HWD \times C)$ are linearly projected into low-dimensional matrices $V_{spatial}(p \times C)$ and $K_{spatial}(p \times C)$, respectively. To facilitate communication between the branches of spatial and channel attention, the weights of the query and key mapping functions are shared, achieving Paired-Attention. This operation also reduces the total number of network parameters. Finally, the spatial attention map and channel attention map are fused through convolutional operations:

$$X = Conv_1(Conv_3(A_s + A_c)). \tag{3}$$

The $Conv3$ represents a convolutional block with a $3 \times 3 \times 3$ kernel size, while $Conv1$ represents a convolutional block with a $1 \times 1 \times 1$ kernel size.

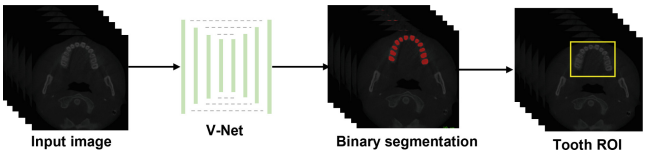

Fig. 3. Computing the ROI of teeth from 3D CBCT image.

3.2 Obtaining ROI of Teeth

The first step for the input 3D CBCT image is to obtain the Region of Interest (ROI) containing teeth. This step can reduce the computational workload for the subsequent tooth centroid and skeleton extraction phase, as well as the segmentation phase. Moreover, it has the potential to improving the overall segmentation accuracy. The specific pipeline of this step is illustrated in Fig. 3. V-Net is used to perform binary segmentation of the image (without distinguishing individual teeth), resulting in the tooth's foreground region. Then, the tooth ROI can be computed from this foreground region.

In order to accurately compute the ROI, the loss function used for training in this step is the combination of the Dice loss and the Cross-Entropy loss.

$$L_{s1} = L_{seg} = \alpha \cdot L_{dice} + (1 - \alpha) \cdot L_{ce}, \tag{4}$$

where

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i q_i + \epsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} q_i^2 + \epsilon}, \tag{5}$$

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^{N} \left(q_i \log\left(p_i\right) + \left(1 - q_i\right) \log\left(1 - p_i\right) \right). \tag{6}$$

Here, p_i represents the value of the i-th voxel in the predicted result, q_i represents the value of the i-th voxel in the ground truth label, and ϵ is a very small number used to prevent division by zero.

3.3 Extraction of Teeth Centroids and Skeletons

The tooth centroid helps determine the tooth's position and instantiate its label, while the tooth skeleton provides an approximate representation of the tooth's morphological structure. By combining the centroid and skeleton information, they can provide guidance for the tooth instance segmentation. The process of this step is illustrated in Fig. 4.

The 3D image is processed by two UNETR++ sub-networks, each containing two decoders. One decoder predicts the binary segmentation map, while the other predicts the 3D offset map. The centroid offset map represents the offset between each voxel and its corresponding tooth centroid, while the skeleton offset is the offset between each voxel and the nearest point on the tooth skeleton.

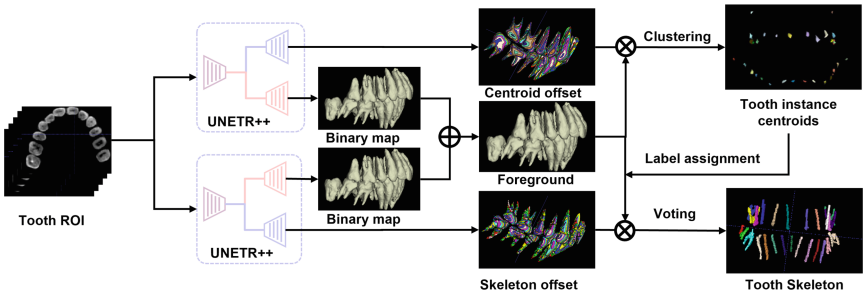

Fig. 4. Extract the centroid and skeleton of teeth

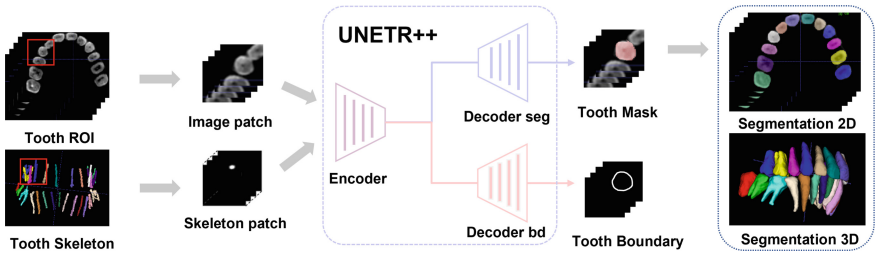

Fig. 5. Complete instance segmentation of teeth

By adding the tooth centroid offset vector to the current foreground voxel coordinates, the tooth centroid density map is obtained. After obtaining the tooth centroid density map, a clustering method [27] is applied to cluster the tooth centroid density map to get tooth instance centroid labels. These labels are then mapped onto the tooth foreground, resulting in instance-level tooth foreground images. Similarly, by using the tooth foreground and skeleton offset vector maps together in the clustering operation, the final instance-level teeth skeleton labels are obtained.

In this step, the loss function considers both the tooth mask segmentation and the tooth centroid or skeleton parts,

$$L_{s2} = L_{seg} + L_{cs}, \tag{7}$$

where L_{seg} represents the loss for tooth mask segmentation, which combines Dice loss and Cross-Entropy loss. L_{cs} represents the loss for tooth centroid or skeleton, using L1 Loss.

3.4 Tooth Instance Segmentation

The final step for tooth instance segmentation is illustrated in Fig. 5. After obtaining the tooth ROI and tooth skeleton, each individual tooth can be cropped around its centroid. The cropped tooth, along with its skeleton, is then concatenated and used as the input to the multi-task UNETR++ model. The

model's output simultaneously predicts tooth masks and tooth boundaries, aiming to maintain accurate tooth segmentation while minimizing errors in tooth surface segmentation.

The loss function for individual tooth segmentation considers both tooth mask segmentation (L_{seg}) and tooth boundary segmentation (L_b):

$$L_{s3} = \lambda L_{seg} + \mu L_b. \tag{8}$$

Here, L_{seg} represents the loss for tooth mask segmentation, which combines Dice loss and Cross-Entropy loss. L_b represents the loss for tooth boundary segmentation, using L2 Loss. In the experiments, $\lambda = 0.6$ and $\mu = 0.1$.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate the performance of our method on the tooth dataset from [9]. This dataset consists of 100 three-dimensional CBCT images of teeth. After excluding two cases where the tooth images did not match the corresponding annotation labels, we were left with 98 valid data cases. Throughout the experiments, the complete dataset was randomly split into 70 cases for training, 8 samples for validation, and 20 samples for testing.

Data Preprocessing. First, we normalize each CBCT image to the range [0, 1]. The specific data preprocessing at different stages is as follows. (1) Obtaining tooth ROI: Due to the limitations of GPU memory, the input tooth CBCT images are randomly cropped to a size of $256 \times 256 \times 256$. (2) Extracting tooth centroids and skeletons: The tooth centroid uses the center of the tooth label. The distance-transform-based algorithm is used to obtain the tooth skeleton [23], which iteratively removes voxels from the binary mask until the skeleton is extracted. After that, the tooth image, tooth label, and tooth skeleton are randomly cropped to a size of $128 \times 128 \times 128$ for training. (3) Tooth instance segmentation: The tooth boundaries are computed by the Canny edge detection algorithm [3] on each 2D CT slice of the 3D CBCT image.

Implementation Details. The experiments were conducted using the PyTorch framework and a GeForce RTX 3090 GPU. The batch sizes for the three stages are 1, 1, and 4. The initial learning rates are set to 0.001, 0.001, and 0.0001 for the three stages. A polynomial learning rate decay strategy is used, which continuously decreases the learning rate during training. The Adam optimizer with a weight decay of 0.0001 is used for optimization. The number of iterations for the three stages are set as 30k, 60k, and 50k, respectively.

Evaluation Metrics. In this study, multiple metrics are used to assess the accuracy and surface error of the segmentation model, including Dice similarity coefficient (DSC), Jaccard Index, Average Surface Distance (ASD), Sensitivity (Sen), and Hausdorff distance (HD).

Table 1. Comparison of segmentation accuracy with other models.

Models	DSC(%)	Jaccard(%)	ASD (mm)	HD(mm)
MWTNet [7]	89.6 ± 1.2	82.5 ± 1.7	0.36 ± 0.14	4.82 ± 1.68
ToothNet [11]	91.9 ± 1.3	84.2 ± 1.8	0.30 ± 0.11	2.82 ± 1.02
CGDNet [31]	93.9 ± 0.9	89.2 ± 0.7	0.27 ± 0.03	1.99 ± 0.78
HMGNet [12]	94.8 ± 0.4	89.1 ± 0.9	0.18 ± 0.02	1.52 ± 0.28
Ours	$\mathbf{95.1 \pm 0.3}$	$\mathbf{90.8 \pm 0.5}$	$\mathbf{0.13 \pm 0.02}$	$\mathbf{1.39 \pm 0.24}$

4.2 Experimental Results

Comparison with Other Methods. In order to validate the segmentation performance of the proposed fully automatic tooth segmentation model based on UNETR++, we conducted comparison experiments with state-of-the-art (SOTA) methods for tooth instance segmentation based on CNN. These methods include MWTNet [7] based on 3D-FCN, ToothNet [11] based on 3D RPN, CGDNet [31] which utilizes tooth center guidance and DenseASPP-UNet, as well as HMGNet [12] based on tooth center and skeleton guidance, and V-Net.

In Table 1, we can see that our model can achieve a Dice Similarity Coefficient (DSC) of 95.1% and a Jaccard index (Jaccard) of 90.8%. These values are higher than other comparison methods, which indicates that our model performs the best in terms of accuracy for tooth segmentation. Additionally, our model also shows the smallest values of Average Surface Distance (ASD) and Hausdorff Distance (HD), which are 0.14 mm and 1.39 mm, respectively. These results prove that the proposed fully automatic tooth instance segmentation method based on UNETR++ not only exhibits higher overall similarity in tooth segmentation, but also performs better in tooth edge segmentation.

To summarize, the tooth segmentation model based on UNETR++ proposed in this study outperforms existing tooth instance segmentation models. This is attributed to the introduction of Self-Attention, which allows the model to easily capture the global information from 3D CBCT images. Additionally, tooth centroids and tooth skeletons contribute to the coarse description of teeth and help improve the accuracy of our model. And the high segmentation accuracy achieved for both tooth masks and tooth edges proves that multi-task learning has shown significant effectiveness.

Ablation Experiment. To better investigate the effectiveness of the proposed UNETR++-based tooth instance segmentation model, we conducted ablation experiments while keeping all other experimental conditions the same. In these experiments, the main backbone networks for all three stages were replaced with V-Net as a baseline, and then UNETR++ was used in the second and third stages and compared with the original model. The results are shown in Table 2.

Table 2 shows that using UNETR++ in both the second and third stages led to improvements in the model's performance. Replacing the network for tooth

Table 2. The performance of replacing the UNETR++ with V-Net.

Models	DSC(%)	Jaccard(%)	ASD(mm)	Sen(%)
VNet(s1,s2,s3)	93.7	88.2	0.21	95.8
VNet(s1,s3) + UNETR++(s2)	93.9	88.6	0.18	**96.6**
VNet(s1,s2) + UNETR++(s3)	94.8	90.1	0.13	94.9
VNet(s1) + UNETR++(s2,s3)(Proposed)	**95.1**	**90.8**	**0.13**	96.1

centroid and skeleton extraction with multi-task UNETR++ in the second stage resulted in a 0.2% increase in DSC and a 0.3 mm reduction in ASD. In the third stage, replacing the network for tooth segmentation with multi-task UNETR++ led to a 1.1% increase in DSC and a 0.8 mm reduction in ASD. Lastly, replacing the backbone networks for both the second and third stages with multi-task UNETR++ resulted in a 1.4% increase in DSC and a 0.8 mm reduction in ASD.

These results indicate that the utilization of the UNETR++ network with Self-Attention mechanism in this study significantly improved the tooth segmentation performance compared to the V-Net based CNN tooth segmentation model. This is because in 3D CBCT images, the morphology and positions of teeth are quite complex, and the images themselves contain massive amounts of information. Therefore, obtaining the global information of the images can help improve the model's performance. The introduction of the Self-Attention mechanism in UNETR++ allows for the effective capture of the global information from 3D CBCT images. The EPA (Efficient Pairwise Attention) blocks play a key role in this process.

5 Conclusion and Future Work

In this paper, we investigate a tooth segmentation method for 3D CBCT images. We use UNETR++ as the backbone network due to its low parameter count, low computational requirements, and state-of-the-art performance in medical image segmentation. We establish a fully automated tooth instance segmentation approach. It begins by obtaining the tooth's Region of Interest (ROI) using V-Net. Subsequently, it represents the tooth's morphological structure coarsely by predicting tooth centroids and tooth skeletons. These centroids and skeletons are then used to guide the tooth instance segmentation process. As a result, the fully automated tooth instance segmentation method developed in this paper outperforms other tooth segmentation methods in terms of both tooth instance segmentation accuracy and tooth boundary segmentation error on CBCT images. This also indicates that using a network with Self-Attention mechanisms can achieve excellent segmentation results in tooth segmentation.

Although our method can outperform other comparison methods, some limitations still exist in this research: (1) Despite UNETR++ being a lightweight model with fewer parameters and computational requirements, the training time is still longer than that of simple CNN networks. (2) The model requires multiple steps to provide guidance for the final segmentation, which necessitates training

multiple networks independently for each step. Future work can be focused on the following directions: (1) Researching even lighter segmentation networks or performing data preprocessing to speed up the training process. (2) Refining the model to combine tasks such as obtaining tooth ROI, tooth centroids, tooth skeletons, and tooth masks into a single stage, achieving tooth instance segmentation in a one-stage process. By addressing these limitations and exploring new approaches, the tooth segmentation method can be further improved and applied more effectively in clinical settings.

References

1. Akhoondali, H., Zoroofi, R., Shirani, G.: Rapid automatic segmentation and visualization of teeth in CT-scan data. J. Appl. Sci. **9**(11), 2031–2044 (2009)
2. Barone, S., Paoli, A., Razionale, A.V.: Ct segmentation of dental shapes by anatomy-driven reformation imaging and b-spline modelling. Int. J. Numer. Method. Biomed. Eng. **32**(6), e02747 (2016)
3. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-8**(6), 679–698 (1986)
4. Cao, H., et al.: Swin-UNet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25066-8_9
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
6. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, Y., et al.: Automatic segmentation of individual tooth in dental CBCT images from tooth surface map by a multi-task FCN. IEEE Access **8**, 97296–97309 (2020)
8. Chung, M., et al.: Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation. Comput. Biol. Med. **120**, 103720 (2020)
9. Cui, Z., et al.: A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nat. Commun. **13**(1), 2096 (2022)
10. Cui, Z., et al.: Tsegnet: an efficient and accurate tooth segmentation network on 3D dental model. Med. Image Anal. **69**, 101949 (2021)
11. Cui, Z., Li, C., Wang, W.: Toothnet: automatic tooth instance segmentation and identification from cone beam CT images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6368–6377 (2019)
12. Cui, Z., et al.: Hierarchical morphology-guided tooth instance segmentation from CBCT images. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 150–162. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_12
13. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Gan, Y., Xia, Z., Xiong, J., Zhao, Q., Hu, Y., Zhang, J.: Toward accurate tooth segmentation from computed tomography images using a hybrid level set model. Med. Phys. **42**(1), 14–27 (2015)

15. Gao, H., Chae, O.: Individual tooth segmentation from CT images using level set method with shape and intensity prior. Pattern Recogn. **43**(7), 2406–2417 (2010)
16. 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: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-08999-2_22
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
18. Hiew, L., Ong, S., Foong, K.W., Weng, C.: Tooth segmentation from cone-beam CT using graph cut. In: Proceedings of the Second APSIPA Annual Summit and Conference, pp. 272–275. ASC, Singapore (2010)
19. Jang, T.J., Kim, K.C., Cho, H.C., Seo, J.K.: A fully automated method for 3D individual tooth identification and segmentation in dental CBCT. IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6562–6568 (2021)
20. Ji, D.X., Ong, S.H., Foong, K.W.C.: A level-set based approach for anterior teeth segmentation in cone beam computed tomography images. Comput. Biol. Med. **50**, 116–128 (2014)
21. Keustermans, J., Vandermeulen, D., Suetens, P.: Integrating statistical shape models into a graph cut framework for tooth segmentation. In: Wang, F., Shen, D., Yan, P., Suzuki, K. (eds.) MLMI 2012. LNCS, vol. 7588, pp. 242–249. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35428-1_30
22. Lahoud, P., et al.: Artificial intelligence for fast and accurate 3-dimensional tooth segmentation on cone-beam computed tomography. J. Endod. **47**(5), 827–835 (2021)
23. Lee, T.C., Kashyap, R.L., Chu, C.N.: Building skeleton models via 3-D medial surface axis thinning algorithms. CVGIP Graph. Models Image Process. **56**(6), 462–478 (1994)
24. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
27. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
28. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: UNETR++: delving into efficient and accurate 3D medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
29. Strbac, G.D., Schnappauf, A., Giannis, K., Bertl, M.H., Moritz, A., Ulm, C.: Guided autotransplantation of teeth: a novel method using virtually planned 3-dimensional templates. J. Endod. **42**(12), 1844–1850 (2016)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
31. 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)

32. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. Adv. Neural. Inf. Process. Syst. **34**, 12077–12090 (2021)
33. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
34. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)