Automatic Segmentation of Internal Tooth Structure from CBCT Images Using Hierarchical Deep Learning

SaeHyun Kim[1], In-Seok Song[2], and Seung Jun Baek[1(✉)]

[1] Korea University, Seoul, South Korea
{vnv73,sjbaek}@korea.ac.kr
[2] Korea University Anam Hospital, Seoul, South Korea
densis@korea.ac.kr

Abstract. Accurate segmentation of teeth is crucial for effective treatment planning. Previous approaches attempted to segment a tooth as a whole, which has limitations because most treatments involve internal structures of teeth. In this paper, we propose fully automated segmentation of internal tooth structure, including enamel, dentin, and pulp, which is the first attempt to the best of our knowledge. The task is challenging, because a total of 96 classes of tooth structures need to be identified from a CBCT image. We design a 3-stage process of coarse-to-fine segmentation of tooth structures without compromising the original resolution. We propose Dual-Hierarchy U-Net (DHU-Net) in order to capture hierarchical structures of teeth, and to effectively fuse encoder and decoder features from higher and lower hierarchies. Experiments demonstrate that our method outperforms state-of-the-art methods in both tasks of segmenting the whole tooth and internal tooth structure.

Keywords: Tooth segmentation · Cone-Beam Computed Tomography (CBCT) · 3D Deep Learning · Attention

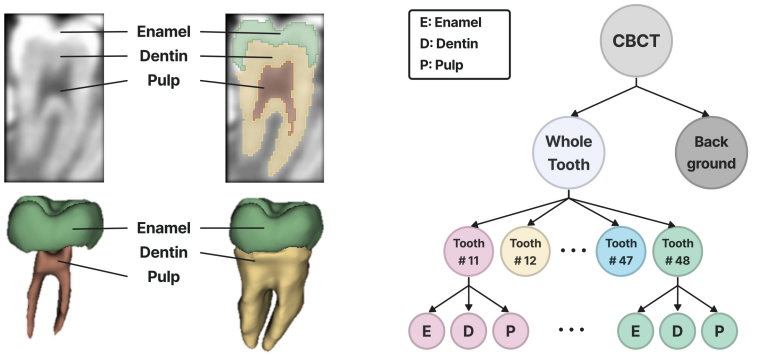
1 Introduction

Cone Beam Computed Tomography (CBCT) is widely used in dental clinics, as it provides volumetric views of tooth structures for diagnosis, treatment, and surgery. Despite the extensive research on teeth segmentation from CBCT images [6,18], segmenting an individual tooth as a whole has limited applications, e.g., predicting tooth movement in orthodontics. Most dental treatments, including caries, prosthodontics and endodontics, focus on the *internal* structures of teeth. Thus, the task of segmenting and representing internal tooth structure is important, and can better assist dental diagnosis and treatment planning [4].

In this paper, we propose an end-to-end framework for tooth segmentation including internal structures from CBCT which, to the best of our knowledge,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43898-1_67.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 703–713, 2023.
https://doi.org/10.1007/978-3-031-43898-1_67



(a) Internal Tooth Structure (b) Hierarchical Structure of Tooth

Fig. 1. Internal Structure of Tooth and its Hierarchy

is the first work to do so. As shown in Fig. 1(a), a tooth consists of *enamel*, *dentin* and *pulp* which we refer to the internal structure. Since there are a total of 32 tooth classes (including wisdom teeth), the model should be capable of identify 96 tooth classes from a CBCT voxel. Considering the size of CBCT data, the problem poses significant challenges from the perspective of not only segmentation performance, but also computational complexity.

We take a hierarchical approach to tackle the challenges, and propose a 3-stage process in order to accurately extract structures without compromising the original resolution of CBCT data. Each stage performs precise detection and segmentation for each level of hierarchy in the tooth structure. We propose a novel module called Dual-Hierarchy U-Net (DHU-Net) which is designed to extract and combine hierarchical features so as to effectively leverage hierarchy in the internal tooth structure. The segmentation performance of our model is evaluated for internal structures as well as the whole teeth. Experiments show that our method outperforms state-of-the-art (SOTA) baselines in both cases.

Our contributions are summarized as follows: 1) a fully automated, end-to-end model for internal tooth segmentation for the first time; 2) a novel 3-stage method with Dual-Hierarchy U-Net module leveraging the hierarchical structures of teeth; 3) the superiority of our model over SOTA baselines.

Related Work. 3D tooth segmentation has been actively studied, including knowledge-based approaches, e.g., graph cut [10] and level set methods [7,8,25] which rely on intensity discrepancies between tooth and non-tooth regions. However, these methods can suffer at regions where teeth meet or where intensity values of roots are similar to jawbone. Internal tooth segmentation methods have been proposed, e.g., enamel-dentin segmentation based on watershed algorithm [13], or tooth pulp cavity segmentation [12,22], which however are sensitive to intensity thresholds, and do not simultaneously segment the entire structure.

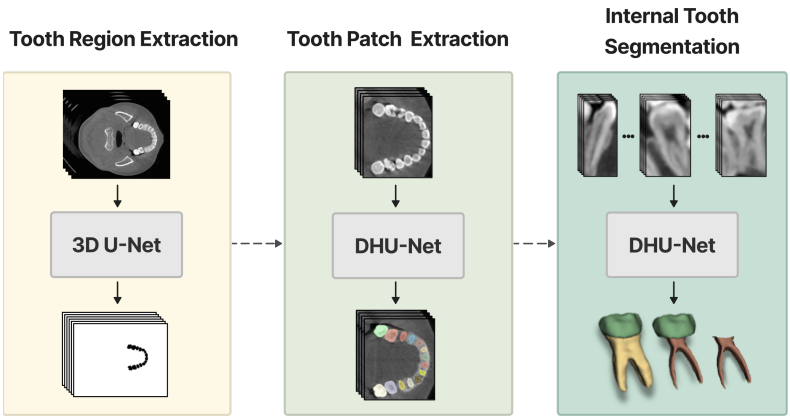


Fig. 2. Three-Stage Process of Internal Tooth Segmentation

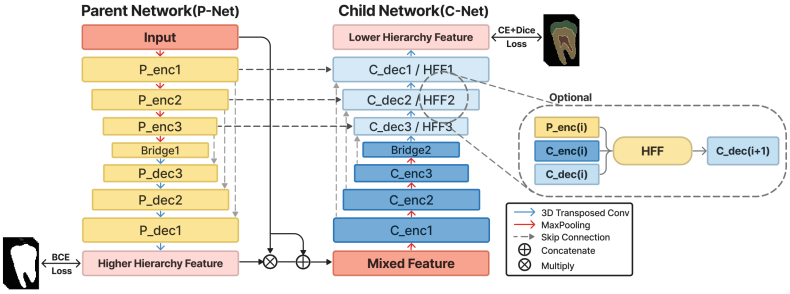
Recently, fully automated segmentation based on deep-learning has been actively explored. ToothNet [3] performs fully automated tooth segmentation using Mask R-CNN [9] which however had limitations, e.g., applicable only to down-sampled CBCT images. Coarse-to-fine segmentation was proposed [5,19], which initially down-sampled and subsequently the full-resolution CBCT images process. SGANet [16] used semantic graph attention based on Graph Convolutional Network [21] to learn and exploit the spatial association among teeth. Prior two-stage approaches [5,14,16,19] extract tooth patches in the 1st stage, and segment the tooth ROIs in the 2nd stage. However, such approaches not only are insufficient for internal segmentation, but also focus on segmenting the individual tooth as a whole, not internal structures.

2 Method

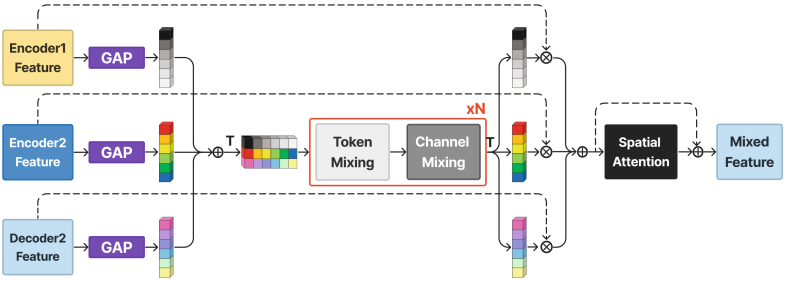
2.1 Three-Stage Segmentation Process

We propose a 3-stage process for the internal tooth segmentation from CBCT images, as shown in Fig. 2. Teeth are categorized into 32 classes of incisors, canines, premolars and molars. Each tooth consists of *enamel*, *dentin* and *pulp*. The union of enamel, dentin and pulp is called the *whole tooth*. Our method performs coarse to fine segmentation based on the following three levels of hierarchy of tooth structures: see Fig. 1(b). (1) a CBCT voxel is classified into tooth and non-tooth; (2) teeth is categorized into 32 classes; (3) a whole tooth is classified into enamel, dentin, and pulp.

Each stage performs the task associated with each level of hierarchy. In Stage 1, a bounding box containing the set of teeth is extracted from CBCT. In Stage 2, 3D patches of individual tooth in 32 classes are extracted. In Stage 3, a tooth patch is segmented into enamel, dentin and pulp structures.



(a) Dual-Hierarchy U-Net (DHU-Net)



(b) Hierarchical Feature Fusion (HFF) Module

Fig. 3. Model Architecture

Stage 1: Tooth Region Extraction. The goal is to extract a 3D bounding box containing the entire set of teeth from CBCT. By removing unnecessary information outside the tooth region, the detection error of tooth can be reduced. We perform a *binary segmentation* of teeth (versus non-tooth) instead of a simple bounding box regression, considering the importance of extracting accurate bounding boxes. After segmentation, we find a tight bounding box around the teeth set which is then zero-padded for extra margins.

We use 3D U-Net [2] for the segmentation of the CBCT image temporarily down-sampled to $128 \times 128 \times 128$ for computational efficiency. Previous methods also proposed to isolate the tooth region, e.g., heuristic thresholding based on Maximum Intensity Projection of CBCT [1]. Our approach may demand more resources, but leads to improved performance, which we show by experiments.

Stage 2: Tooth Patch Extraction. Individual tooth patches are extracted from the tooth region received from Stage 1. A *patch* is a 3D bounding box around an individual tooth. To extract patches, a *segmentation* of tooth into 32 classes is performed. Similar to Stage 1, the purpose of segmentation is precise extraction of patches. The individual tooth patch is created by padding the segmented tooth into size $64 \times 64 \times 96$. We propose Dual-Hierarchy U-Net (DHU-Net) for precise segmentation which leverages hierarchical properties of

tooth features aiming at accurate identification and extraction of ROIs. Detailed architecture of DHU-Net is described in Sect. 2.2.

Stage 3: Internal Tooth Segmentation. The individual tooth patch is segmented into enamel, dentin and pulp. DHU-Net is again used in Stage 3 for a precise segmentation, which is explained in Sect. 2.2.

Design Insights. Our method prunes non-tooth regions from the CBCT in Stage 1, and extracts tight tooth patches in Stage 2. The process reduces false-negatives (missed detection of teeth), and false-positives (segmenting irrelevant regions), promoting a precise segmentation of internal structures in Stage 3.

2.2 Dual-Hierarchy U-Net (DHU-Net)

Dual-Hierarchy U-Net (DHU-Net) consists of two cascaded U-Net which are called Parent Network (P-Net) and Child Network (C-Net) as shown in Fig. 3(a). P-Net (resp. C-Net) learns features of the higher (resp. lower) level of hierarchy. Importantly, both P-Net and C-Net output segmentation maps which are supervised by the labels of corresponding hierarchy:

- Stage 2: The P-Net output is supervised by binary (tooth and non-tooth) labels. The C-Net output is supervised by 32-class teeth labels.
- Stage 3: The P-Net output is supervised by binary (whole tooth and background) labels. The C-Net output is supervised by the labels of internal structures.

The output from P-Net promotes improved segmentation in C-Net as follows. Let I_p and I_c denote the inputs to P-Net and C-Net respectively, and Z_p denote the output feature of P-Net, respectively. The input of C-Net, I_c, is defined as $I_c = I_p \oplus (I_p \otimes \sigma(Z_p))$ where $\oplus$, $\otimes$ and σ represent concatenation, element-wise multiplication and sigmoid function respectively. $I_p \otimes \sigma(Z_p)$ is the output from P-Net *gated* by the segmentation map. Thus, C-Net receives the input with the highlighted ROIs (the entire teeth in Stage 2 or a whole tooth in Stage 3) in addition to the raw input, which facilitates the fine-level segmentation at C-Net.

In addition, the decoder layers of C-Net utilize Hierarchical Feature Fusion (HFF) module for effective fusion of the features from P-Net and C-Net, as explained in the next section. DHU-Net is inspired by double U-Net [11], however differs from it in several ways: the supervision of P-Net and C-Net outputs with labels at high- and low-level hierarchies, the way input and output of P-Net are combined, and the existence of HFF module.

2.3 Hierarchical Feature Fusion (HFF) Module

One of the properties that made U-Net successful is the combination of encoder and decoder features through skip connections. In the proposed Hierarchical Feature Fusion (HFF) module, the decoder layers at C-Net combines *two* encoder

Table 1. Comparison of Internal Tooth Segmentation

	Enamel		Dentin		Pulp		
Method	DSC	HD(95%)	DSC	HD(95%)	DSC	HD(95%)	DP(%)
2-Stage							
3D UNet	79.08 ± 0.81	2.74 ± 0.38	82.84 ± 0.85	2.17 ± 0.33	75.91 ± 1.13	2.81 ± 0.29	93.64
Att UNet	81.74 ± 1.20	2.11 ± 0.32	84.51 ± 1.10	1.91 ± 0.26	75.54 ± 1.04	2.57 ± 0.34	94.79
3-Stage							
3D UNet	83.36 ± 0.62	1.56 ± 0.24	86.42 ± 0.32	1.61 ± 0.12	77.02 ± 0.66	2.58 ± 0.11	96.53
Att UNet	83.66 ± 0.24	1.53 ± 0.17	86.00 ± 0.43	1.71 ± 0.19	77.35 ± 0.51	2.49 ± 0.13	97.68
Ours	$\mathbf{85.65 \pm 0.29}$	$\mathbf{1.37 \pm 0.26}$	$\mathbf{88.05 \pm 0.31}$	$\mathbf{1.45 \pm 0.16}$	$\mathbf{78.58 \pm 0.38}$	$\mathbf{2.12 \pm 0.12}$	**98.84**

features from both hierarchies, i.e., P-Net and C-Net: see Fig. 3(a). HFF facilitates the propagation of hierarchical features over the network.

As shown in Fig. 3(b), the concept of Channel Attention [23,24] is used in HFF. Attention vectors are created by mixing pooled features using MLPMixer [20]. The feature maps are scaled in a channel-wise manner by the attention vectors, and then fused after applying spatial attention. The overall process allows the model to effectively highlight important channel and spatial features from multiple hierarchies.

2.4 Loss Function

The loss function of DHU-Net is given by

$$L_{\text{total}} = L_{\text{P}} + \lambda_1 \cdot L_{\text{C}} + \lambda_2 \cdot L_{\text{FTM}} \tag{1}$$

L_{P} and L_{C} are binary cross-entropy (CE) loss for P-Net and CE + DICE loss for C-Net, respectively. λ_1 and λ_2 are hyperparameters for balancing losses. The λ_1 and λ_2 are hyperparameters for balancing losses which are set to 2 and 5, respectively. L_{FTM} is Focal Tree-Min Loss [15], a hierarchical loss function encouraging the model to capture hierarchical relationships between the features extracted by P-Net and C-Net, e.g., the features of a whole tooth and its internal structures in Stage 3.

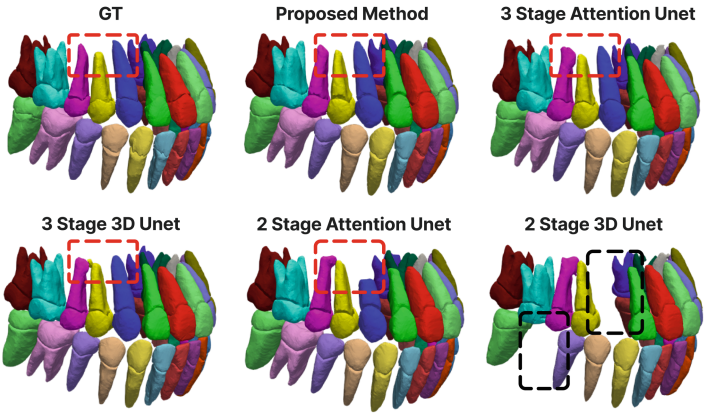
3 Experiment

3.1 Dataset

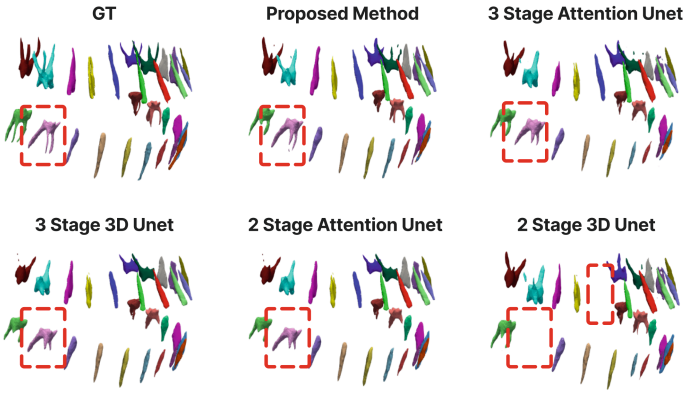
The dataset consisted of 70 anonymized cases of 3D dental CBCT images collected from the Korea University Anam Hospital. This study was approved by the Institutional Review Board of of the same hospital (IRB No. 2020AN0410). The dimension of CBCT images is $768 \times 768 \times 576$ with the voxel size $0.3 \times 0.3 \times 0.3\,\text{mm}^3$. We clipped the intensity values of CBCT images to $[-1000, 2500]$ and applied intensity normalization.

Table 2. Comparison of the Whole Tooth Segmentation

Method	DSC	Jaccard	HD95
C2FSeg [5]	88.68 ± 1.43	80.59 ± 2.09	3.12 ± 1.45
MWTNet [1]	90.18 ± 1.04	82.62 ± 1.16	2.78 ± 1.38
SGANet [16]	92.16 ± 0.45	86.48 ± 0.78	2.24 ± 0.54
Ours	$\mathbf{93.91 \pm 0.34}$	$\mathbf{88.67 \pm 0.68}$	$\mathbf{1.32 \pm 0.30}$



(a) Visualization of Internal Tooth Segmentation (Enamel, Dentin)



(b) Visualization of Internal Tooth Segmentation (Pulp)

Fig. 4. Qualitative Analysis of Internal Tooth Segmentation

All the internal structures of teeth in CBCT images were individually labelled as enamel, dentin, and pulp. The labeling was performed by two experts and cross-checked, with a final inspection performed by a oral & maxillofacial surgeon. The dataset is split in 3:1:1 for train, validation and test with 5-fold nested cross-validation. We use the following metrics: Dice similarity coefficient (DSC), Jaccard index, and Hausdorff distance (HD95). We evaluate the accuracy of tooth identification (in 32 classes) during the patch extraction in Stage 2. We define the metric of detection precision (DP) as $\mathrm{DP} = |D \cap G|/|D \cup G|$, where D represents the set of predicted tooth classes in Stage 2, and G represents the ground truth set. All the results are averaged over 10 repetitions of experiments.

3.2 Experimental Results

We evaluated the performance of our model for internal tooth segmentation by comparing with two commonly used models in medical segmentation: U-Net [2] and Attention U-Net [17]. We consider the cases of two- and three-stage process for baselines. For two stages, baselines perform extraction of tooth patches from the CBCT image in the 1st stage, and internal tooth segmentation from the patch in the 2nd stage. The three-stage process is identical to our model, except that the segmentation networks are replaced by the baseline models.

Table 1 shows the segmentation performance of internal tooth structures. Our method outperforms the baselines across all the metrics (comparison of Jaccard is provided in Supplementary Materials). By comparing 2-stage and 3-stage processes for baselines, we observe that the 3-stage process leads to the better performance. This shows the importance of reducing detection errors, i.e., accurate extraction of tooth patches in turn enhances the final segmentation performance. Indeed, 3-stage process improves the DP metric in all cases. In addition, by comparing with 3-stage baselines, we observe that DHU-Net outperforms U-Net and Attention U-Net. The results demonstrate the effectiveness of the hierarchical design of deep learning models for analyzing internal tooth structure. Ablation analysis on some components of DHU-Net, i.e., HFF module and hierarchical loss function, is provided in Supplementary Materials.

We conducted a qualitative analysis of segmentation results as shown in Fig. 4. We found that the U-Net baseline had a problem of missed detection of teeth, while the 2-stage Attention U-Net showed a cut-out problem. Our model resulted in better representations of the root parts of dentin and pulp compared to the 3-stage baselines, perhaps because our model was better at dealing with the problem of similar intensity values of teeth and the jaw bone.

Next, we evaluate the segmentation performance of the *whole tooth*, which also is an important problem. Our model provides the prediction of the whole tooth, i.e., we can simply take a union of the predicted enamel, dentin and pulp. We selected state-of-the-art methods for tooth segmentation as baselines: C2FSeg [5], MWTNet [1] and SGANet [16]. As shown in Table 2, our approach outperformed the baselines, and proved to be effective for segmenting the whole tooth as well.

We observe that by comparing Table 1 and 2, the segmentation performance of whole tooth is higher than that of internal structures. This is reasonable, because the segmentation of finer structures tend to be harder. For example, suppose our model incorrectly classified an enamel voxel as dentin. This does not affect the accuracy of the whole tooth prediction, however, the accuracy of *both* enamel and dentin predictions will drop in the internal segmentation task.

4 Conclusion

In this work, we proposed a fully automated segmentation of internal tooth structures, which, to the best of our knowledge, is the first attempt. We proposed a 3-stage process to reduce detection error and overcome difficulties in segmentation and computational complexity. We introduced DHU-Net, a segmentation network capable of effectively learning hierarchical features of tooth structures, demonstrating improved segmentation performance for both the whole tooth and internal structures. Our future work include the segmentation of additional structures from CBCT, such as mandible or maxilla, simultaneously with teeth.

Acknowledgements. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2020-0-01819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1007215 and No.2022R1A5A1027646), and the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711195279 , RS-2021-KD000009).

References

1. 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)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
3. 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)
4. Ezhov, M., et al.: Clinically applicable artificial intelligence system for dental diagnosis with cbct. Sci. Rep. **11**(1), 15006 (2021)
5. Ezhov, M., Zakirov, A., Gusarev, M.: Coarse-to-fine volumetric segmentation of teeth in cone-beam ct. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 52–56. IEEE (2019)
6. Gan, Y., Xia, Z., Xiong, J., Li, G., Zhao, Q.: Tooth and alveolar bone segmentation from dental computed tomography images. IEEE J. Biomed. Health Inf. **22**(1), 196–204 (2017)

7. 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)
8. 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)
9. 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)
10. 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)
11. Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D.: Doubleu-net: a deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 558–564. IEEE (2020)
12. Jiang, B., et al.: Dental pulp segmentation from cone-beam computed tomography images. In: The Fourth International Symposium on Image Computing and Digital Medicine, pp. 80–85 (2020)
13. Kakehbaraei, S., Seyedarabi, H., Zenouz, A.T.: Dental segmentation in cone-beam computed tomography images using watershed and morphology operators. J. Med. Signals Sensors **8**(2), 119 (2018)
14. Lee, J., Chung, M., Lee, M., Shin, Y.G.: Tooth instance segmentation from cone-beam ct images through point-based detection and gaussian disentanglement. Multimedia Tools Appl. **81**(13), 18327–18342 (2022)
15. Li, L., Zhou, T., Wang, W., Li, J., Yang, Y.: Deep hierarchical semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1246–1257 (2022)
16. Li, P., et al.: Semantic graph attention with explicit anatomical association modeling for tooth segmentation from cbct images. IEEE Trans. Med. Imaging **41**(11), 3116–3127 (2022)
17. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
18. Rao, Y., Wang, Y., Meng, F., Pu, J., Sun, J., Wang, Q.: A symmetric fully convolutional residual network with dcrf for accurate tooth segmentation. IEEE Access **8**, 92028–92038 (2020)
19. Shaheen, E., et al.: A novel deep learning system for multi-class tooth segmentation and classification on cone beam computed tomography: a validation study. J. Dentistry **115**, 103865 (2021)
20. Tolstikhin, I.O., et al.: Mlp-mixer: an all-mlp architecture for vision. Adv. Neural Inf. Process. Syst. **34**, 24261–24272 (2021)
21. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
22. Wang, L., Li, J.p., Ge, Z.p., Li, G.: Cbct image based segmentation method for tooth pulp cavity region extraction. Dentomaxillofacial Radiol. **48**(2), 20180236 (2019)
23. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)

24. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
25. Xia, Z., Gan, Y., Chang, L., Xiong, J., Zhao, Q.: Individual tooth segmentation from ct images scanned with contacts of maxillary and mandible teeth. Comput. Methods Prog. Biomed. **138**, 1–12 (2017)