DAE-Net:Dual attention embedding-based tooth instance segmentation approach for panoramic X-ray images

Liangyu Chen[1,‡], Zheng Li[1,‡], Dongping Zhang[1], Tianxu Yan[1⋆], Yutong Wei[2], and Luying Qian[1]

[1] College of Information Engineering, China Jiliang University, Hangzhou, China
[2] College of Computer and Control Engineering, Qiqihar University, Qiqihar, China

Abstract. Tooth instance segmentation is a key technology in the field of medical image segmentation, with applications ranging from orthodontic treatment to dental pathology assessment. Although researchers have developed many tooth instance segmentation models, a common drawback is that traditional techniques are not effective in combining global and local background information and fail to make full use of labeled data, resulting in limited performance in different clinical scenarios. This paper proposes a new deep learning method dedicated to the instance segmentation task of dental panoramic images. By introducing the RFEM attention module and the CBAM attention module into the ResNet architecture, we improve the network's ability to focus on key features by adaptively assigning weights to different channels and further enhance the fusion effect of global and local features. In order to improve the generalization ability of the data, we use a data augmentation algorithm specifically for labeled datasets. Finally, the network is named the fully supervised tooth segmentation model DAE-Net, which performs well on the MICCAI STS 2024 dataset, with a final performance improvement of 32% compared to the original base model, verifying its effectiveness and stability in the tooth instance segmentation task.

Keywords: Panoramic X-Ray images, Instance Segmentation, Dual Attention Mechanism, ResNet

1 Introduction

Dental imaging involves the segmentation of teeth, maxillary and mandibular bones, and surrounding tissues to capture detailed images that are critical to doctors' diagnosis, treatment selection, and auxiliary treatment. The importance of panoramic imaging in dental diagnosis has become increasingly prominent. Panoramic X-rays are not only a means of visualizing various diseases and abnormalities but also a basic method of dental imaging. It helps doctors assess

⋆ Corresponding author: 1594842227@qq.com. First Author and Second Author contribute equally to this work and should be regarded as co-first authors.

dental health, detect cavities, identify impacted teeth, and analyze the alignment of a patient's teeth. However, currently, dentists usually need to manually interpret panoramic X-rays and analyze the shape, number, and growth position of teeth, which is a time-consuming and labor-intensive process [6].

With the development of artificial intelligence, computer vision and deep learning technology have made significant progress in automatic panoramic X-ray analysis, aiming to assist dentists in accurate diagnosis and significantly improve the diagnostic efficiency and accuracy of panoramic X-rays. Convolutional neural networks (CNN) have changed the landscape of medical image analysis by automatically learning and extracting hierarchical features in images and are widely used in image classification, segmentation, and detection tasks, especially in breast tumors, lung lesions, and heart disease. Important breakthroughs have been made in detection [7,2,10]. In recent years, deep learning has also made progress in the field of dental image segmentation, and Faster R-CNN and Mask R-CNN are widely used in tooth segmentation, although these methods rely on pre-positioned teeth[13,1]. MSLPNet solves the boundary prediction problem and achieves precise positioning of teeth by introducing multi-scale structures[3]. In addition, TransUNet combines U-Net and Transformer architectures to improve image segmentation performance[9].

However, these methods often suffer from the problems of low instance segmentation accuracy and large types of dental images that are difficult to distinguish. This paper addresses this issue and makes the following contributions:

- We proposed a tooth instance segmentation network based on a dual attention mechanism, aiming to improve the accuracy of instance segmentation of dental panoramic images.
- We proposed a tooth instance segmentation network based on a dual attention mechanism, aiming to improve the accuracy of instance segmentation of dental panoramic images.
- We proposed a tooth instance segmentation network based on a dual attention mechanism, aiming to improve the accuracy of instance segmentation of dental panoramic images.

2 Method

In this paper, we will introduce a new dual attention embedding method for tooth instance segmentation, namely DAE-Net, to solve the problem of low accuracy and difficulty in distinguishing tooth image instance segmentation. The workflow of DAE-Net is divided into two main stages: In the first stage, the region proposal network generates candidate boxes of potential objects, classifies these regions, fine-tunes the bounding boxes, and generates pixel-level masks. The second stage is the mask generation stage, in which the network scans the proposed regions generated in the first stage and creates object categories, bounding boxes, and masks for each region. We will describe the specific details of the network in the following chapters.

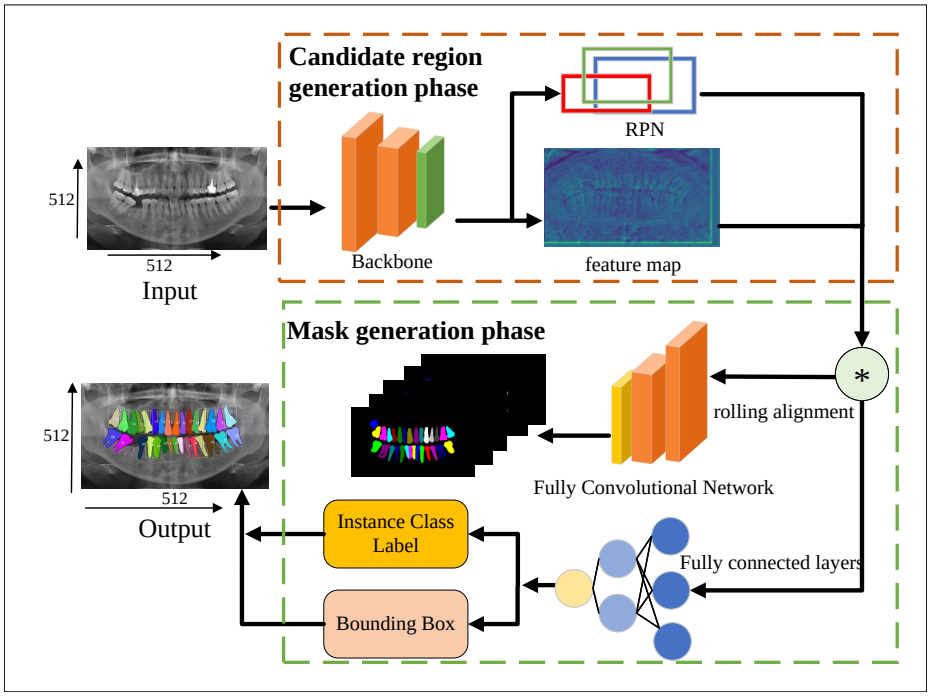

Fig. 1: Overview of the DAE-Net network structure, which mainly includes the region proposal generation stage and the mask generation stage.

2.1 Data Preprocessing

The images are uniformly scaled from the original 1991×1127 to the preset 512×512. In addition, we use translation, horizontal flipping, vertical flipping, and diagonal flipping to expand the data.

2.2 Proposed Method

In this paper, we propose a dual attention embedding-based tooth instance segmentation network, namely DAE-Net. The network mainly consists of a region proposal generation stage and a mask generation stage. In the region proposal generation stage, the region proposal network scans the feature map in a top-down path of the feature pyramid to extract possible object locations. In the mask generation stage, the regions are processed by different neural networks to generate object categories, bounding boxes, and pixel-level masks. In particular, in the region proposal generation stage, we embed RFEM and CBAM in the backbone. RFEM enhances important features by adaptively adjusting channel weights. CBAM combines channel and spatial attention mechanisms to further optimize feature representation capabilities. Below we will introduce the two key components in detail.

2.3 Refinement Feature Enhancement Module

In order to enhance the network's ability to capture key information, we designed a refined feature enhancement module, which compresses effective features by means of feature refinement and promotes the stability of feature distribution through the weights of the linear layer. The detailed structure is shown in Figure 2. Specifically, the feature F is first compressed by an adaptive average pooling operation, which aggregates features of size H×W to obtain an aggregated information feature map. In this process, the global distribution of channel features is saved and used as the weight of different channel features. Then the tensor shape is adjusted to remove redundant information and improve the module's operating efficiency.

After the linear layer compresses the dimension, it passes through the activation function, activates specific channel features through the channel weight allocation mechanism, controls the activation of each channel, and then multiplies the feature F with the corresponding channel weight to obtain the enhanced result FRF. The specific calculation formula is as follows:

$$F_c = \text{linear}_1\left(R_{s_1}\left(\text{Avg}(F)\right)\right) \tag{1}$$

$$F_c' = \text{linear}_2\left(\text{relu}\left(F_c\right)\right) \tag{2}$$

$$F_c^s = R_{s_2}\left(\text{Sigmoid}\left(F_c'\right)\right) \tag{3}$$

$$F_{RF} = F \times F_c^s \tag{4}$$

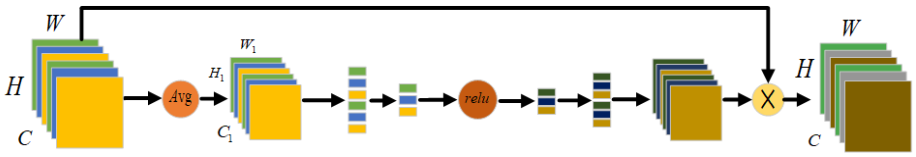

Fig. 2: RFEM modules.

2.4 Convolutional Block Attention module

In order to optimize the feature representation within each residual block, important features are further highlighted. We embed the CBAM module in the network, which contains two submodules, namely the channel attention module (CAM) and the spatial attention module (SAM). The detailed structure is shown in Figure 3. Specifically, the feature map $S \in \mathbb{R}^{C\times H\times W}$ is input into the CBAM, and the one-dimensional channel attention module $M_c \in \mathbb{R}^{C\times 1\times 1}$ is input into the CBAM, and the one-dimensional channel attention module $M_s \in \mathbb{R}^{1\times H\times W}$ are generated in turn. In this process, the attention value is replicated in the spatial dimension, and finally the optimization resultS'' is output. The specific calculation can be seen in formulas 5.

$$S' = M_c(S) \otimes S \quad , \quad S'' = M_s(S') \otimes S' \tag{5}$$

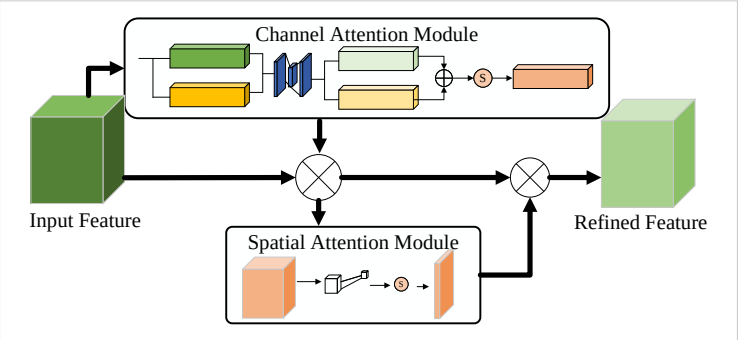

Fig. 3: CBAM modules.

In CAM, the feature map S compresses the spatial dimension through global average pooling and global maximum pooling to obtain two feature vectors $S^{c}_{avg}$and S^{c}_{max}then passes through the fully connected layer, ReLU activation function, fully connected layer, and element-by-element addition operation to obtain the channel attention map, which is then passed through the sigmoid function and multiplied element-by-element with the feature map S to finally obtain Mc. In SAM, the input feature map is subjected to global average pooling and global maximum pooling to generate two two-dimensional feature maps, Sc_avg and Sc_max, which are spliced through the channel dimension to form a more effective feature descriptor to highlight the important spatial area. Finally, a 7×7 convolutional layer is used to generate the spatial attention map Ms. The above process is expressed as formulas (6, 7).

$$\begin{aligned}\mathbf{M_c}(\mathbf{F}) &= \sigma\left(\mathrm{MLP}(\mathrm{AvgPool}(S)) + \mathrm{MLP}(\mathrm{MaxPool}(S))\right) \\ &= \sigma\left(\mathbf{W}_1\left(\mathbf{W}_0\left(S^{\mathbf{c}}_{\mathrm{arg}}\right)\right) + \mathbf{W}_1\left(\mathbf{W}_0\left(S^{\mathbf{c}}_{\mathrm{max}}\right)\right)\right)\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbf{M_s}(\mathbf{F}) &= \sigma\left(f^{7\times7}([\mathrm{AvgPool}(S); \mathrm{MaxPool}(S)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[S^{c}_{\mathrm{arg}}; S^{c}_{\mathrm{max}}\right]\right)\right)\end{aligned} \tag{7}$$

2.5 Post-processing

Since the setting of the IoU threshold will affect the prediction output of the model, it is necessary to set the IoU threshold reasonably. After experiments, we found that when the IoU threshold is adjusted to 0.2–0.5, the accuracy and reliability of the model prediction results can be effectively improved. Figures 4 and 5 show the results before and after the IoU threshold is adjusted.

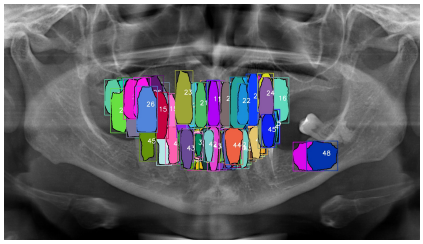

Fig. 4: Before IoU threshold adjustment.

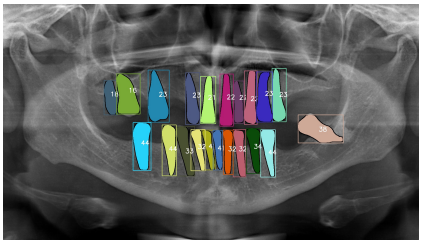

Fig. 5: After IoU threshold adjustment.

2.6 Loss Function

In this study, in order to balance the model's focus on region overlap and edge accuracy, we use a weighted loss function that combines Dice and IoU to train the model. The specific formula is as follows:

$$\text{loss} = \text{DeepDiceLoss}(S_{\text{deep}}, Y_{\text{deep}}) \times 0.6 + \text{DeepIoULoss}(\hat{Y}_{\text{deep}}, Y_{\text{deep}}) \times 0.4 \quad (8)$$

3 Experiments

3.1 Datasets and Comparative Models

The dataset used in this experiment comes from 30 labeled tooth images[12,5,4] provided by the MICCAI STS 2024: Panoramic X-ray Images Challenge. Since the amount of labeled data is too small, we use effective data augmentation methods to expand the data, and finally obtain 600 images. In addition, this study will divide the expanded data into 70%, 10%, and 20%. ratios, and adjust the size of the input data to 512×512. In order to evaluate the effectiveness and superiority of this method, we compared it with the methods of the other three contestants and compared it with two advanced medical image instance segmentation networks(U-Net[8] and SegFormer[11]) in the instance segmentation network part.

3.2 Evaluation metrics

To evaluate the performance of the model in the experiment, we used the official evaluation indicators of the MICCAI 2024 challenge, namely DSC_instance, DSC_image, NSD_instance, NSD_image, mIoU_instance, mIoU_image, and IA.

3.3 Implementation details

Environment Setup and Training Scheme. In this study, our detailed configuration and training hyperparameters are shown in Table 1. In addition, it should be emphasized that we did not use any unlabeled data in the entire training process.

Table 1: Development environments and requirements

Environment Setup		**Training Program**	
System	Windows 11	Batch size	32
CPU	i5-12600KF CPU @4.9GHz	Train size	512 × 512
RAM	16GB	Total number of iterations	200000
GPU version	NVIDIA GeForce RTX 4060 TI 8G	Optimizer	Adam
CUDA version	11.0	Initial learning rate (lr)	0.0001
Programming language	Python3.8	Training time	40 hours
Deep learning framework	torch1.11.0		

3.4 Ablation Study

In this section, we perform ablation experiments on key components, aiming to reveal their specific contributions. Furthermore, we choose MaskRcnn as the backbone. The specific ablation results are shown in Table 2.

REFM of Effectiveness. By comparing the data in the second and third rows of Table 2, we find that the performance of the model is significantly improved after integrating REFM in the backbone network. Specifically, Dice_Instance increased from 0.339 to 0.41, and MIoU_image increased from 0.057 to 0.247. These results clearly demonstrate that REFM plays a key role in improving model performance.

CBAM of Effectiveness. By comparing the experimental data in the second and fourth rows of Table 2, we can find that the performance of the model has also been improved after integrating CBAM. Specifically, Dice_Instance increased to 0.40, and MIoU_image increased to 0.566. These experimental results fully demonstrate the importance of CBAM in improving model performance.

CBAM and RFEM of Effectiveness. Observing the data in Table 2, we found that the model achieved the best results in all indicators after combining CBAM and RFE, which fully proves that the combination of CBAM and RFEM has a greater impact on improving model performance. important role.

Table 2: Results of ablation experiments on various DAE-Net components.

Frame	image-level			instance-level			
	↑Dice%	↑IoU %	↑NSD%	↓NSD%	↑IoU %	↑NSD%	↑IA%
Backbone	0.339	0.45	0.089	0.484	0.057	0.292	0.040
Backbone + REFM	0.410	0.733	0.278	0.761	0.247	0.582	0.251
Backbone + CBAM	0.400	0.720	0.203	0.749	0.177	0.566	0.184
Backbone + CBAM + REFM	0.430	0.741	0.274	0.770	0.270	0.595	0.267

4 Results and discussion

DAE-Net is a novel tooth image instance segmentation model designed by us. The experimental results will be further analyzed and discussed below.

4.1 Quantitative results

In order to evaluate the superiority of the proposed method more objectively and fairly, we compared DAE-Net with the methods of the other three contestants in the MICCAI STS 2024 challenge (Contestant 1, Contestant 2, and Contestant 3). In addition, this study also replaced the instance segmentation network in MaskRcnn with Unet and SegFormer for experimental comparison. Table 3 shows the detailed comparison results.

Comprehensively observing various performance indicators, the results show that DAE-Net has achieved significant superiority in all key evaluation indicators.

Table 3: Comparison results with other contestants' models.

Frame	image-level			instance-level			
	↑Dice%	↑IoU %	↑NSD%	↓NSD%	↑IoU %	↑NSD%	↑IA%
Contestant 1	31.6%	78.6%	22.0%	83.8%	15.9%	65.4%	16.5%
Contestant 2	13.5%	70.3%	21.4%	73.8%	16.7%	56.0%	19.5%
Contestant 3	6.2%	71.8%	21.6%	74.6%	18.9%	56.5%	18.5%
Unet	33.9%	45.0%	8.9%	48.4%	5.7%	29.2%	4.0%
SegFormer	35.0%	72.0%	20.3%	74.9%	17.7%	56.6%	18.4%
DAE-Net (Our)	43.0%	74.1%	27.4%	77.0%	24.4%	59.5%	26.7%

4.2 Qualitative results

In order to more intuitively demonstrate the superiority of this method, we visualized the instance segmentation results of three models, including DAE-Net. The specific comparison results are shown in Figure 8.

By observing the visualization results, we can find that DAE-Net has an advantage in the accuracy of tooth instance segmentation. In addition, we found that the three comparison models can complete effective segmentation, and the DAE-Net proposed in this paper achieved the best performance.

4.3 Limitation and future work

This study mainly uses dental panoramic images to design supervised models without considering the impact of label errors of certain categories in the data on the model prediction results and the use of unlabeled data. Therefore, in future research work, we will consider more about developing dental panoramic image instance segmentation methods based on semi-supervised learning or unsupervised learning.

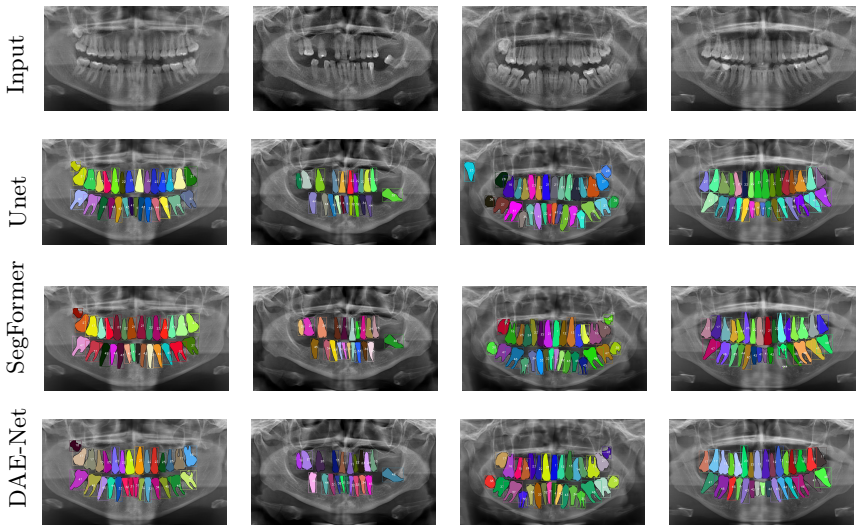

Fig. 6: Visualization of test results.

5 Conclusion

In order to solve the problem of low segmentation accuracy caused by the morphological differences between different categories of teeth, we proposed a tooth instance segmentation method for panoramic X-ray images based on dual attention embedding (DAE-Net). Specifically, DAE-Net is a two-stage segmentation method that includes region proposal generation and mask generation. In the region proposal generation stage, the region proposal network extracts feature maps from the feature pyramid network to generate possible target regions containing objects. In the mask generation stage, we use different neural networks to process the target region, generate object categories, bounding boxes, and pixel-level masks, and use ROIAlign technology to accurately locate the relevant parts in the feature map. However, in order to improve the model's ability to focus on important features and efficiently utilize them, we embed two important components in the backbone, namely RFEM and CBAM. From the results of the above experiments, it can be found that after embedding the two key components, our model has achieved a significant improvement in overall performance, and has also achieved certain results in the latest challenge (MICCAI STS 2024). In general, since we only develop a fully supervised tooth instance segmentation model for a very small number of labeled data, we lack consideration and utilization of unlabeled data. Therefore, in the future we plan to further explore the valuable information implied by unlabeled data in the context of semi-supervision or weak supervision to further develop more practical tooth instance segmentation models.

Acknowledgements

This work was supported by Zhejiang Key R & D Project of China (2024C01102, 2024C01108, 2023C01030, 2022C01082).

References

1. Brahmi, W., Jdey, I.: Automatic tooth instance segmentation and identification from panoramic x-ray images using deep cnn. Multimedia Tools and Applications **83**(18), 55565–55585 (2024) 2
2. Chen, C., Zhou, K., Zha, M., et al.: An effective deep neural network for lung lesions segmentation from covid-19 ct images. IEEE Transactions on Industrial Informatics **17**(9), 6528–6538 (2021) 2
3. Chen, Q., Zhao, Y., Liu, Y., et al.: Mslpnet: multi-scale location perception network for dental panoramic x-ray image segmentation. Neural Computing and Applications **33**, 10277–10291 (2021) 2
4. Cui, W., Wang, Y., Li, Y., et al.: Ctooth+: A large-scale dental cone beam computed tomography dataset and benchmark for tooth volume segmentation. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 64–73 (2022) 6
5. Cui, W., Wang, Y., Zhang, Q., et al.: Ctooth: a fully annotated 3d dataset and benchmark for tooth volume segmentation on cone beam computed tomography images. In: International Conference on Intelligent Robotics and Applications. pp. 191–200 (2022) 6
6. Hingst, V., Weber, M.A.: Dental x-ray diagnostics with the orthopantomography–technique and typical imaging results. Der Radiologe **60**, 77–92 (2020). https://doi.org/10.1007/s00117-019-00620-1 2
7. Robin, M., John, J., Ravikumar, A.: Breast tumor segmentation using u-net. In: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). pp. 1164–1167 (2021) 2
8. Siddique, N., Paheding, S., Elkin, C.P., et al.: U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access **9**, 82031–82057 (2021) 6
9. Sun, D., Wang, J., Zuo, Z., et al.: Sts-transunet: Semi-supervised tooth segmentation transformer u-net for dental panoramic image. Mathematical Biosciences and Engineering **21**(2), 2366–2384 (2024) 2
10. Wolterink, J.M., Leiner, T., Viergever, M.A., et al.: Dilated convolutional neural networks for cardiovascular mr segmentation in congenital heart disease. In: International Workshop on Reconstruction and Analysis of Moving Body Organs. pp. 95–102 (2016) 2
11. Xie, E., Wang, W., Yu, Z., et al.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021) 6
12. Zhang, Y., Ye, F., Chen, L., et al.: Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data **10**(1), 380 (2023) 6
13. Zhu, Y., Xu, T., Peng, L., et al.: Faster-rcnn based intelligent detection and localization of dental caries. Displays **74**, 102201 (2022) 2