Semi-Supervised 2D dental image segmentation via Cross Teaching network

Lai Wang[1] and Wentao Bao[2]

[1] School of Communication Engineering, Hangzhou Dianzi University, China
[2] School of Automation, Hangzhou Dianzi University, China
241080002@hdu.edu.cn

Abstract. Many deep learning models are capable of performing various image segmentation tasks with excellent performance on well-annotated datasets. 2D panoramic dental image segmentation, as a specific task within the field of medical image segmentation, plays a crucial role in providing reference material for auxiliary diagnosis or subsequent downstream tasks. However, the reality is that there is a scarcity of well-annotated related datasets in this field, making the development of accurate and efficient semi-supervised dental image segmentation methods a considerable challenge. This paper presents a teeth segmentation method based on a semi-supervised segmentation research via cross teaching between CNN and Transformer. And we designed and used some pre-processing and post-processing methods to complete the migration of the original research on the 2D dental panoramic X-ray tooth images segmentation task. Finally, Our method achieved an average instance Dice score of 87.07% and average instance NSD score of 41.08% for the teeth segmentation on the validation set and instance using a NVIDIA GeForce RTX 4090. The average running time was 2.36 seconds for one image. The code is available at https://github.com/aicorein/STS2024-Semi-Supervised-Cross-Teaching.

Keywords: dental image segmentation · semi-supervised learning · cross teaching.

1 introduction

Medical image segmentation is a very important field in deep learningin recent years, which is a core task in the field of medical imaging, aiming to separate different structures or tissues in medical images by training on a large number of fine-labeled or unlabeled medical images. It has made great contributions to precise diagnosis, disease prediction, personalized treatment, assisted surgery, and medical training in medical imaging. After decades of development, from the classic convolutional neural network (CNN) to the U-Net network[7] specially designed for biomedical image segmentation, which adopts an encoding-decoding structure, it has greatly promoted the development of medical image segmentation. However, due to various limitations, it is difficult to obtain

medical datasets, and classical neural network methods require extremely fine voxel-level and pixel-level annotation, which requires a lot of time and effort. Semi-supervised segmentation, a learning method that uses labeled and unlabeled data, has become popular in recent years. Compared with the usual fully supervised methods, semi-supervised segmentation can reduce annotation costs, improve the model's generalization ability, learn more image features, and better handle class imbalance problems.

Before the rise of deep learning, the main medical image processing techniques used were thresholding, edge detection, and region growing, which also gradually revealed their drawbacks: poor handling of image noise and poor adaptability to complex structures. With the development of deep learning, convolutional neural networks have played a huge role, such as FCN[3]that brought the image segmentation field into a new era, and U-Net that brought the medical image segmentation field to new heights. Variants such as U-Net++[12], DeepLab[1], and 3D medical image segmentation using U-Net have also been derived. Semi-supervised medical image segmentation: Semi-supervised learning (SSL) is a branch of deep learning that is suitable for handling mixed data sets of limitedlabeled images and large unlabeled images. With the development of deep learning, training methods such as sub-training,deep co-training[6][11], graphical models, generative adversarial networks[2], and generating pseudo-labels[8] have been developed.

This paper mainly studies the research on 2D dental panoramic X-ray tooth images. 2D panoramic X-ray is an effective method for dentists to determine children's hidden cavities, affected teeth and extra teeth. However, manual segmentationof teeth from panoramic X-rays requires huge human effort and time, and cannot obtain a large number of effective annotated case images. This work transfers specific tasks based on a cross-teaching model of CNN and Transformer[5][4], and designs new pre-processing and post-processing. The original model combines the spatial feature extraction of CNN and the global modeling capability of Transformer, which greatly improves the effective accuracy compared to traditional segmentation. Among them, CNN uses the U-Net model as the backbone to extract features from the input labeled and unlabeled images, and extracts important features of 2D dental X-ray images through convolutional layers and pooling layers to generate feature maps; at the same time, the Transformer part inputs the feature map generated by CNN through the self-attention mechanism to process the complex relationship between features. The advantages of this method are as follows: (1) Simultaneously equipped with the local convolution mechanism of CNNs and the long-term attention mechanism of Transformer, enhancing feature extraction ability. (2) Cross-teaching is an implicit consistency regularization method,we simplify co-training from implicit consistencyregularization to cross-teaching, where one network predicts as pseudo-labels and end-to-end supervises the other network, enabling the model to remain stable under certain perturbations. In this work, we mainly made the following two contributions: (1) Through specific experiments, it is verified that the model can achieve good results in the specific task of tooth segmentation. (2) Designed and

implemented some pre-processing and post-processing methods as part of the segmentation task workflow.

2 Method

We have adapted an existing cross-teaching mechanism based on CNN and Transformer for the current 2D X-ray dental image segmentation task. The specific method is mainly divided into two stages: training and inference. During the training phase, similar to the original model's approach, the training dataset is divided into labeled D_l and unlabeled D_u. The model's dataloader ensures that each batch maintains a certain ratio of images from D_l to those from D_u. Images from D_l in a batch, after passing through the upper and lower main networks, will produce the loss $\mathcal{L}_{sup}$, while images from D_u will produce the loss $\mathcal{L}_{ctl}$ through a special computation method after forward propagation through CNN and Transformer networks. Subsequently, the total loss $\mathcal{L}_{total}$ as weighted sum of $\mathcal{L}_{sup}$ and $\mathcal{L}_{ctl}$ serves as the overall network loss, which is then used to guide gradient descent. In the inference phase, only the CNN network is used for prediction.

2.1 Proposed Method

Based on the experimental results of the original model[5], this model selects the combination with the best experimental results from the original model: the CNN network U-Net with an additional Transformer network Swin-UNet as the backbone of network. Fig. 1 shows the complete network structure of this model. Since there are relatively few labeled data in the training set, the use of unlabeled image data will be an important factor affecting the final performance of the model. Consistent with the methods previously mentioned, The use of unlabeled data is mainly reflected in calculating the loss value $\mathcal{L}_{ctl}$ of the unlabeled image, and then integrated with loss $\mathcal{L}_{sup}$ to perform specific gradient descent operations on both networks. For the components and calculation steps of the loss function, please refer to the paper[5] of the original model. We have not made any changes on this part.

2.2 Preprocessing

Pre-processing The data preprocessing procedure involves converting the original X-ray JPEG images into 8-bit single-channel grayscale images. Additionally, it is necessary to generate label maps that serve as ground truth. For labeled images, the original JSON annotations contain a list of contour coordinates and classification categories for each tooth. During processing, the image is first scaled to 224×224, and all the coordinates in the coordinate list are also scaled proportionally. To enhance the model's generalization capabilities, the contours are not used directly for training. Instead, the pixels occupied by the teeth are filled based on the contour coordinates before being processed by the model. The

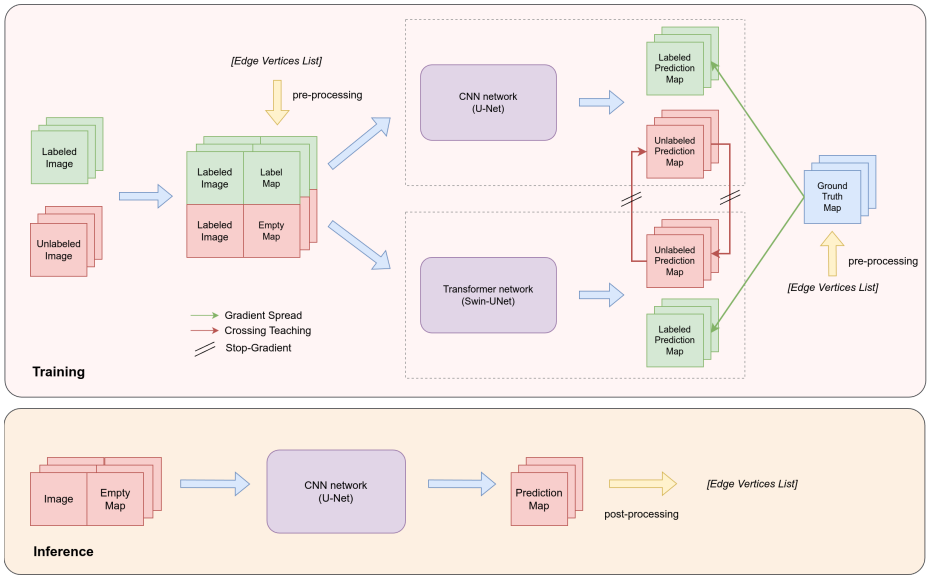

Fig. 1. Network architecture (includes the network structure of training and inference phases)

specific steps are as follows: first, initialize a blank 8-bit single-channel grayscale image (all pixel values start at 0); then map the tooth label value t_l to the classification value t_c ($0 \leq t_c \leq 52$) through a simple label-to-category correspondence table; subsequently, calculate the pixel area occupied by the current tooth using the contour point list and fill the area with classification value. In this way, the label map can be obtained from the JSON annotation of a single image. For unlabeled images, the label map is initialized as an full-zero value grayscale image.

Post-processing The training process almost does not require additional post-processing. Once the loss calculation is completed, gradient descent can be performed. However, the inference process requires post-processing. The results generated by inference are 8-bit single-channel grayscale images (where the pixel values occupied by each tooth correspond to the classification values). Therefore, the classification values are first converted to label values, and then edge extraction operations are performed on the polygonal pixel areas of tooth. A simple algorithm is used here for edge extraction: if a pixel's value is not 0, and the values of the four neighboring pixels are all the same as this pixel's value, then this pixel is judged to be a non-edge pixel of the tooth polygonal area and needs to be updated to a value of 0 in the subsequent process. By traversing every point on the image, in the end, all non-0 value pixels on the image represent the edge points of each tooth, and the values of these pixels are the classification labels of teeth.

3 Experiments

3.1 Dataset

The dataset used for training is provided by the competition[10], and the data belongs to the 2D panoramic X-ray images category. The specific ratio is shown in Table 1. The image size is almost all in the ratio of 16 : 9, but the image width varies from 1000 to 3000 pixels. In addition, the image is a 3-channel JPEG image.

Table 1. Data statistics on training and validation sets.

Data part name	Num of Labeled	Num of Unlabeled
Training	30	2353
Validation	0	20

3.2 Evaluation metrics

Mainly includes these metrics:
(1) Dice Similarity Coefficient (DSC): instance-level and image-level
(2) Normalized Surface Distance (NSD): instance-level and image-level
(3) mean Intersection-over-Union (mIoU): instance-level and image-level
(4) Identification Accuracy (IA)

3.3 Implementation details

Environment settings The development environments and requirements are presented in Table 2. The system is running on WSL2 Ubuntu 24.04 LTS as the operating system. The CPU in use is an Intel(R) Core(R) 13700K CPU with a clock speed of 5.40GHz. The system has a total of 64GB RAM, divided into 4 modules of 16GB each, operating at a speed of 10400MT/s. The system is equipped with one NVIDIA 4090 24G GPU. The CUDA version installed on the system is 12.6. The programming language used for development is Python 3.9.7. The deep learning framework employed includes torch 1.10.2, torchvision 0.11.3 and Pillow 8.4.0.

Training protocols For data augmentation, it is performed in real-time within the dataloader during training, as the augmentation process is not a time-consuming performance bottleneck, thus it is done at runtime to avoid excessive storage space usage. The augmentation methods are relatively simple: each image has a 50% chance of undergoing a random angle rotation transformation, a 50% chance of a 90-degree rotation transformation, and then a random flip

Table 2. Development environments and requirements.

System	Ubuntu 24.04 LTS in Microsoft WSL2
CPU	Intel(R) Core(R) 13700K CPU@5.40GHz
RAM	4×16GB; 10400MT/s
GPU (number and type)	One NVIDIA GeForce RTX 4090 24G
CUDA version	12.6
Programming language	3.9.7
Deep learning framework	torch 1.10.2, torchvision 0.11.3, Pillow 8.4.0

Table 3. Training protocols.

Batch size	$16 \times 1 \times 224 \times 224$
Total epochs	43 (2383 its/epoch)
Optimizer	SGD (momentum $\mu = 0.99$)
Initial learning rate (lr)	0.01
Training time	12 hours
Number of model parameters	7.28M

along the x or y axis. After the transformation, the image still maintains a size of 224×224, but the blank areas are filled with value of 0. Since the aspect ratio of the tooth images in this segmentation task hardly changes, no further scaling transformations are applied to the effective part of the image. For sampling, each training batch is generated by a two-stream data loader, which ensures that the ratio of labeled and unlabeled data in each batch is always $1 : 1$.

4 Results and discussion

4.1 Quantitative results on validation set

The metrics scores on the validation set show in Table 4. The low IA score is somewhat related to the non-smooth segmentation edges. Since the IA score evaluation requires the IOU area of the corresponding image to be greater than a certain value, the non-smooth edges actually result in a smaller IOU value, which is more likely to be considered as unsuccessful segmentation.

4.2 Qualitative results on validation set

Good segmentation cases In case 04 and case 12 in Fig. 2. The cases with better segmentation show better image features. These samples are almost all

Table 4. Quantitative evaluation results.

Method	image-level			instance-level			
	Dice (%)	IoU (%)	NSD (%)	Dice (%)	IoU (%)	NSD (%)	IA (%)
Final model	37.66	58.17	72.78	87.07	33.00	41.08	4.78

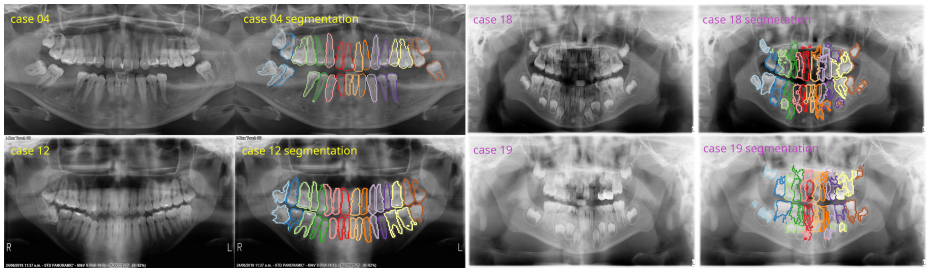

Fig. 2. Good and failure segmentation cases: good cases are 04 and 12, failure cases are 18 and 19.

images of adults' teeth. These teeth have good symmetry and are evenly distributed. In addition, the images of these cases have good contrast, and the teeth and background are clearly distinguishable, which is also a good basis for segmentation.

Failure case analysis In case 18 and case 19 in Fig. 2. Failed cases are often accompanied by broken, messy segmentation edges, and multiple discontinuous areas are identified as the same tooth. Most of these examples are images of children's teeth. Children also go through the process of tooth replacement, so there may be multiple teeth in the vertical direction at the same horizontal axis position. Moreover, the morphological characteristics of deciduous teeth and permanent teeth are different, which also brings challenges to correct segmentation. And in some samples, such as the image of sample 19, the difference between the tooth area and the background is not obvious, which will also cause interference.

4.3 Ablation Studies

We also conducted ablation studies to explore the impact of unlabeled image data on model performance. The specific method involved: eliminating the cross teaching loss from the loss for unlabeled data. Additionally, only labeled image data was utilized for training. Training was halted when the model achieved the same Dice score as the complete model during epoch evaluation. As competitors, we do not have access to the ground truth of the validation set, hence we cannot obtain metrics scores on the validation set. However, a simple analysis can be conducted by examining specific instance segmentation results. The comparison of two cases is shown in the Fig. 3. It can be clearly seen that if the unlabeled

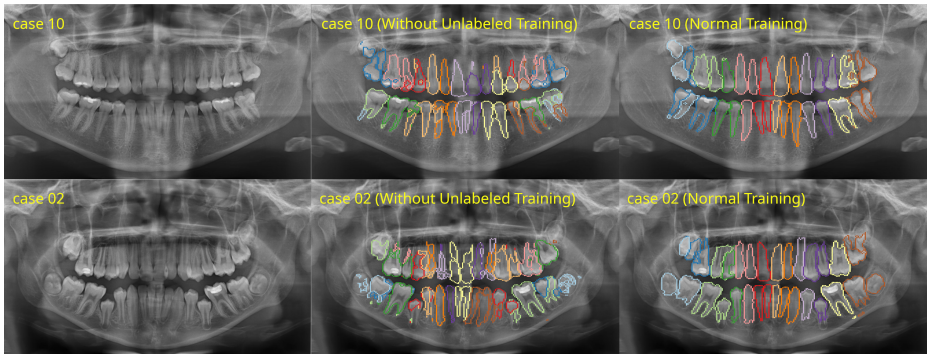

Fig. 3. Comparsion between training without labeled and training normally.

image data is utilized, the understanding of tooth shape and semantics between different teeth can be effectively improved during segmentation, avoiding the formation of broken segmentation edges.

4.4 Results on final testing set

We obtained scores of 59.68% (image-level Dice), 43.01% (image-level IoU), 86.96% (image-level NSD), 38.54% (instance-level Dice), 34.71% (instance-level IoU), 72.52% (instance-level NSD) and 8.6% IA on the official test set. The total time latency on the test set was 11.8102 seconds and average time latency on the validation set was 2.36 seconds, with an average area under the GPU memory-time curve of 13910.32. Collectively, we ranked eight among all submitted teams.

4.5 Limitation and future work

Due to the characteristics of the network backbone, larger input images cannot be utilized, hence the necessity to downscale images during both training and inference. This downscaling affects the acquisition and understanding of valid information within the images during training, and also impacts the precision of the classification mask output during inference. As a result, after the post-processing stage, it leads to the generation of non-smooth segmentation edges, which significantly impacts the IA metric in evaluation. Future work will focus on employing superior network models, which will enhance the network's overall performance on one hand, and on the other hand, attempt to address the issue of non-smooth segmentation edges. This could involve using more sophisticated polygon edge fitting or smoothing algorithms. Alternatively, the approach could be improved by training outputting the key vertices of the tooth polygon and redrawing the edges directly on higher-resolution images.

5 Conclusion

In conclusion, this work introduces a semi-supervised model based on cross-teaching between CNN and Transformer into the segmentation task of 2D X-ray panoramic dental images, and employs some pre-processing and post-processing to handle the input and output images. The experimental results indicate that although the model has issues such as lower evaluation metric scores and rough segmentation edges, it occupies relatively fewer resources and has acceptable inference speed. Also, its output can be used for basic reference and further analysis. In the future, we will continue to improve and develop better network architecture and operational workflows.

Acknowledgements. The authors of this paper declare that the segmentation method they implemented for participation in the STS 2024 challenge has not used any additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention. We thank all the data owners for making the X-ray images and CT scans publicly available and Codebench [9] for hosting the challenge platform.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017) 2
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf 2
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015) 2
4. Luo, X.: SSL4MIS. https://github.com/HiLab-git/SSL4MIS (2020) 2
5. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International Conference on Medical Imaging with Deep Learning. pp. 820–833. PMLR (2022) 2, 3
6. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Proceedings of the european conference on computer vision (eccv). pp. 135–152 (2018) 2
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015) 1

8. Wang, G., Zhai, S., Lasio, G., Zhang, B., Yi, B., Chen, S., Macvittie, T.J., Metaxas, D., Zhou, J., Zhang, S.: Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. IEEE transactions on medical imaging **41**(3), 531–542 (2021) 2
9. Xu, Z., Escalera, S., Pavao, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7) (2022) 9
10. Zhang, Y., Ye, F., Chen, L., et al.: Childrens dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data **10**(1), 380 (2023) 5
11. Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A.: Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 121–140. IEEE (2019) 2
12. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018) 2

Table 5. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	2
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	4
Pre-processing	3
Strategies to use the partial label	4
Strategies to use the unlabeled images.	3
Strategies to improve model inference	4
Post-processing	4
The dataset and evaluation metric section are presented	7
Environment setting table is provided	6
Training protocol table is provided	6
Ablation study	7
Efficiency evaluation results are provided	7
Visualized segmentation example is provided	7
Limitation and future work are presented	Yes
Reference format is consistent.	Yes