Deformable Inherent Consistent Learning Network for Accurate Tooth Segmentation in Dental Panoramic Radiographs

Xinxu Cai[0009−0002−7681−9947], Yisong Zhang[0009−0004−7205−2640], Zeyuan Guan[0009−0006−1422−6408], Qi Sun[0009−0001−9135−8408], and Zhenshen Qu*[0000−0002−2436−9932]

Harbin Institute of Technology: Harbin, CN
{23s004026,23s004040,23s004079,23s121074}@stu.hit.edu.cn, ocicq@126.com

Abstract. This study presents a Deformable Inherent Consistent Learning (DICL) network for tooth segmentation in dental panoramic radiographs, addressing the clinical need for accurate diagnostic tools amid the complexity of dental diseases. The DICL network enhances segmentation accuracy by effectively learning from a limited labeled dataset, guiding robust categorical representation of teeth. It employs deformable convolution to capture detailed dental features and a two-stage training strategy, transitioning from semi-supervised to fully supervised learning, optimizing performance with labeled and pseudo-labeled images. Achieving scores of 84.45% (image-level Dice), 73.60% (image-level IoU), 88.57% (image-level NSD), 22.36% (instance-level Dice), 57.39% (instance-level IoU), 69.85% (instance-level NSD) and 65.82% IA on the official test set, our method demonstrates superior segmentation effectiveness, especially in boundary and root segmentation, outperforming algorithms like U-Net. This study's contributions advance dental image segmentation, improving diagnostic efficiency and reducing errors. The code is available at https://github.com/Dew026/DICL.git.

Keywords: Dental Image Segmentation · Deep Learning · Deformable Convolution · Inherent Consistency .

1 Introduction

In the realm of oral health, dental diseases are a category of conditions that severely impact human quality of life. Not only do they lead to serious pathological changes within the oral cavity, but they can also adversely affect other organs throughout the body, thereby affecting an individual's overall health status [3]. Currently, in the clinical treatment of dental diseases, there is a heavy reliance on the clinical experience of dentists, panoramic X-ray films, and consultation diagnoses to formulate treatment plans. However, as clinical practice continues to deepen, an increasing number of patients are found to suffer from multiple dental diseases simultaneously, which raises the bar for dentists in terms of etiological analysis and planning of treatment strategies. Therefore, the development of an effective deep learning algorithm for tooth segmentation is of great significance for simplifying the diagnostic process for dentists, reducing the rates of misdiagnosis and missed diagnosis, and improving the efficiency of medical work.

In the complex medical environment, the segmentation of dental images often faces various challenges, including variability in image quality and the complexity of tooth structures. These challenges not only impair diagnostic accuracy but can also lead to serious treatment risks [6]. Therefore, accurate tooth segmentation detection has become a critical step in ensuring the safety of the diagnostic process and the quality of treatment. Currently, although rule-based methods combined with manual visual inspection are widely used in dental image segmentation, these methods are often inefficient and susceptible to subjective factors, leading to missed or false detections.

In recent years, the novel applications of deep learning technology have revolutionized the field of dental image segmentation. Its powerful feature extraction capabilities enable the automatic extraction of features from dental images and the identification of tooth structures, providing precise segmentation boundaries. For instance, Panfilov et al. [5] applied Mixup image augmentation technology during the training process and validated its effectiveness in dental image segmentation, thereby enhancing the robustness of the model . Chaitanya et al. [1] proposed a learning-based generative network that learns from both labeled and unlabeled data simultaneously and applies real spatial deformation fields and additive intensity transformation fields to synthesize new samples. In addition, some researchers have incorporated prior domain knowledge as a regularization term into the segmentation model to address semi-supervised segmentation issues [8]. Dong et al. [2]introduced a deep graph network that includes a lightweight registration network and multi-level information consistency constraints. He et al. [4] proposed a method that uses an autoencoder to extract multi-scale semantic features from unlabeled data to assist the segmentation network.

In summary, despite significant progress in the field of automatic tooth segmentation, there is still an urgent need to develop a method that comprehensively considers the variability of image quality, the complexity of tooth structures, and segmentation accuracy. Such a method would be able to automatically recognize tooth structures, improve segmentation efficiency and quality, reduce labor and time costs, and thus bring a breakthrough in the field of dental image segmentation.

In this study, we selected a dataset of dental panoramic radiographic images from the STS 2024 Challenge. The main contributions of this work are as follows:

1. A Deformable Inherent Consistent Learning Network is proposed, which obtains tooth-focused feature information through deformable convolution and further learns the inherent consistency between labeled and unlabeled images using the network.

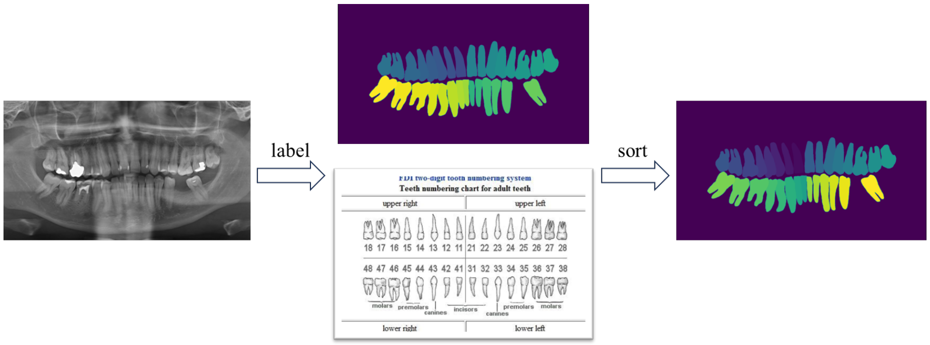

Fig. 1. The process of generating new label.

2. A two-stage training network framework is proposed, which inherits the weights from the semi-supervised learning network into the fully supervised learning network, and further trains and optimizes the results through labeled images and enhanced pseudo-labeled images.

2 Method

2.1 Preprocessing

In the context of dental image data for this competition, due to the lack of color information in X-ray images, we have converted all images to grayscale to facilitate subsequent training computations. Additionally, considering the large size of the original images and the variability in dimensions across different image types, we have employed random cropping to a fixed patch size, standardizing the image size for convenient network training.

Furthermore, the original images utilize an internationally recognized dental numbering system for annotating each tooth, which does not commence with a sequential format starting from zero. To enhance the network's predictive capabilities, we have mapped these annotations to a compact sequential format starting from zero, which is amenable for classification as shown in Fig. 1.

Taking into account that all feature values are on the same scale, gradient descent algorithms can more swiftly locate the optimal solution. We have normalized the grayscale values of the original images using min-max normalization to stabilize numerical computations and accelerate the convergence rate of gradient descent algorithms within neural networks.

2.2 Proposed Method

For semi-supervised tasks, the training set typically comprises a small labeled dataset $D_l = \{(x_i^l, y_i^l)\}_{i=1}^n$ and a much larger unlabeled dataset $D_u = \{x_j^u\}_{j=1}^m$(where

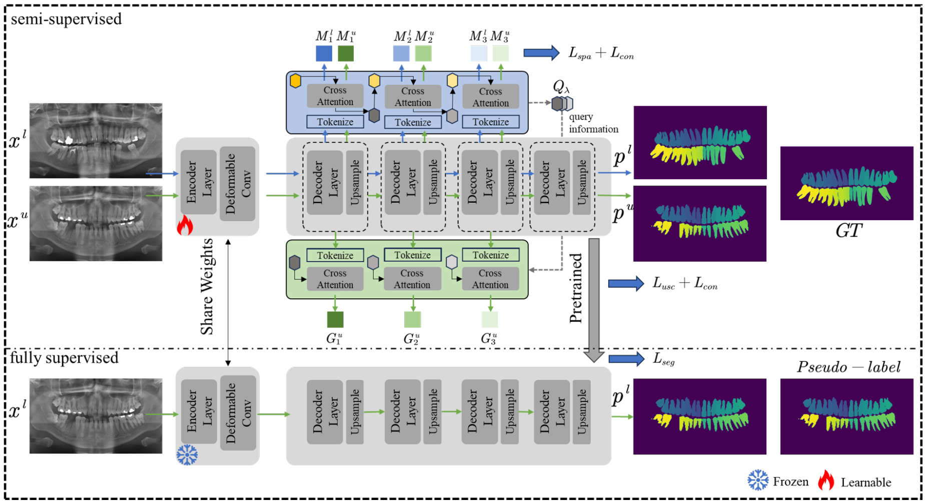

Fig. 2. The network structure of proposed DICL. In the semi-supervised phase, DICL employs consistent learning with both labeled and unlabeled images, predicting pseudo labels for the unlabeled ones, which are then fed into the fully supervised part. During the fully supervised phase, we forecast the segmentation mask for unlabeled tooth images and iteratively refine the pseudo labels.

$n \ll m$). In this work, we propose a two-stage Deformable Inherent Consistent Network aimed at effectively learning robust categorical representations on $D_l \cup D_u$ under the guidance of a limited number of D_l, thereby enhancing segmentation accuracy. Specifically, we introduce a deformable convolution module based on the Inherent Consistent Network, allowing the network to initially focus the feature map more on dental information.

As illustrated in Fig. 2 , our network architecture consists of two training phases: semi-supervised and . During the semi-supervised training phase, in the image encoding stage, we use deformable convolution to further focus the feature information on individual teeth, adaptively learning the relevant shape characteristics of the teeth.

In the image decoding stage, we adopt a structure similar to the Inherent Consistent Network, which mainly includes the following steps:

1. On the output feature map, by introducing initial query information, labeled and unlabeled features are used to generate the next-scale query information Q_λ and predictions for each scale through the cross-attention mechanism M_λ^l and M_λ^u.
2. For unlabeled feature information, we directly use the query information Q_λ generated in the previous step through the cross-attention mechanism to produce unlabeled predictions for each scale.

During the training process of step (1), we consider the loss function between the predictions at each scale M_λ^l and the annotated ground truth y^l.

$$\mathcal{L}_{\mathrm{spa}} = \frac{1}{\lambda} \sum_{\lambda} \left[\mathcal{L}_{dice} \left(\sigma \left(I \left(M_\lambda^l \right) \right), y^l \right) + \mathcal{L}_{ce} \left(\sigma \left(I \left(M_\lambda^l \right) \right), y^l \right) \right] \tag{1}$$

where $\lambda \in (1, 2, 3)$ represents different scales, $\sigma(\cdot)$denotes the softmax operation, and $I(\cdot)$ is the bilinear interpolation to upsample the predictions to the same size as label y^l.

In step (2), we first consider the loss between the unlabeled predictions G_λ^u generated at each scale and the final predictions p^u generated by the network.

$$\mathcal{L}_{\mathrm{usc}} = \frac{1}{\lambda} \sum_{\lambda} \left[\mathcal{L}_{dice} \left(\sigma \left(I \left(G_\lambda^u \right) \right), \sigma \left(p^u \right) \right) \right] \tag{2}$$

Additionally, we consider the semantic consistency between the labeled and unlabeled predictions M_λ^u, thus employing a consistency loss.

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{\lambda} \sum_{\lambda} \left[\mathcal{L}_{MSE} \left(\sigma \left(G_\lambda^u \right), \sigma \left(M_\lambda^u \right) \right) \right] \tag{3}$$

In the fully supervised training phase, the encoding phase of the entire network inherits the weights from the semi-supervised training phase, while the decoding phase employs a basic simple decoder to simplify the model structure as much as possible. In this part, we use the labeled data from the training set and other dental data predicted by the aforementioned semi-supervised training as input, obtaining the final model in a fully supervised manner. The goal of our deformable inherent consistent learning (DICL) framework is to minimize the following combined objective function that contains the supervised and unsupervised parts:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \mathcal{L}_{spa} + \mathcal{L}_{usc} + \mathcal{L}_{total} \tag{4}$$

$$\mathcal{L}_{seg} = \mathcal{L}_{dice} \left(p^l, y^l \right) + \mathcal{L}_{ce} \left(p^l, y^l \right) \tag{5}$$

2.3 Post-processing

When using our network to obtain the final results, there may be small connected regions that are not the desired predicted results and need to be removed. As mentioned earlier, in our approach, the weighted combination of the output masks from the fully supervised part and the semi-supervised part can help address this issue to some extent.

3 Experiments

3.1 Dataset

For this study, we use the dataset from the STS-2024 Challenge. The competition includes a dental panoramic radiographs dataset [10], as shown in Fig. 1. The images in this dataset are of medical nature and are commonly used to assist doctors in diagnosis and treatment. The dataset consists of three-channel tooth X-ray panoramic images with dimensions of 1127×1991 and 942×2000.

The dataset includes a training set, a validation set and a test set. The training set contains 30 labeled tooth images and 2359 unlabeled tooth images. The validation set contains 20 unlabeled tooth images.

3.2 Evaluation metrics

In this study, we use Dice coefficient, IoU (Intersection over Union), and 2D Hausdorff distance as evaluation metrics.

The Dice coefficient is the most commonly used metric in medical image competitions. It is a measure of set similarity and is typically used to calculate the similarity between two samples, which is calculated by:

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{6}$$

where A and B represent the two masks respectively.

The IoU is define as the ratio between area of intersection and area of union, which is calculated by:

$$IoU = \frac{(A \cap B)}{(A \cup B)} \tag{7}$$

where A and B represent the two masks respectively.

The Normalized Surface Distance (NSD) is a boundary-based metric that evaluates the closeness of two segmentation surfaces, often used to measure the accuracy of medical image segmentation algorithms. It is calculated based on the overlap between the boundaries of the segmentation result and the ground truth, within a certain tolerance level. The formula for NSD is given by:

$$\mathrm{NSD}(G, S) = \frac{\|\partial G \cap B_{\partial S}^{(\tau)}\| + \|\partial S \cap B_{\partial G}^{(\tau)}\|}{\|\partial G\| + \|\partial S\|} \tag{8}$$

where G and S represent the ground truth and segmentation result, respectively. ∂G and ∂S denote the boundaries of G and S. $B_{\partial G}^{(\tau)}$ and $B_{\partial S}^{(\tau)}$ are the border regions of the ground truth and segmentation surface at a tolerance τ. The tolerance τ is a user-defined threshold that specifies the maximum allowed distance between the two surfaces for them to be considered as matching.

Identification Accuracy (IA) is a metric used to evaluate the performance of image segmentation algorithms. It measures the proportion of pixels that are correctly identified as belonging to a particular class or not. The IA can be defined using the following formula:

$$IA = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

where True Positives (TP) denote the pixels that have been accurately identified as part of the object or region under consideration. True Negatives (TN) refer to the pixels that have been correctly determined to be outside of the object or region of interest. False Negatives (FN) are those pixels that were mistakenly classified as background or as belonging to another category when they are actually part of the object or region in question. Conversely, False Positives (FP) are pixels that were incorrectly identified as being part of the object or region of interest when they are not.

The IA score ranges from 0 to 1, with 1 indicating perfect identification accuracy where all pixels are correctly classified. This metric is particularly useful for assessing the accuracy of segmentation tasks where the goal is to distinguish between different classes or objects within an image .

3.3 Implementation details

Environment settings The development environment and requirements are listed in Table 1. The operating system running on the system is Windows 11. The CPU used is 12th Gen Intel(R) Core(TM) i9-12900H with a clock speed of 2.50 GHz. The system has a total of 16GB of RAM. It is equipped with an NVIDIA 3060 8G GPU. The system has CUDA version 11.8 installed. The programming language used for development is Python 3.7.4. The deep learning frameworks used include torch 1.13.1 and torchvision 0.14.1.

Table 1. Development environments and requirements.

System	Windows 11
CPU	Intel(R) Core(TM) i9-12900H CPU 2.5GHz
RAM	16GB
GPU (number and type)	One NVIDIA 3060 8G
CUDA version	11.8
Programming language	Python 3.7.4
Deep learning framework	torch 1.13.1, torchvision 0.14.1

Training protocols The strategies for semi-supervised training are shown in Table 2. We used annotated data exclusively in this part. The strategies for fully supervised training are shown in Table 3. In this part, we utilized both annotated and unlabeled data with their pseudo-labels. Taking into account the displacements incurred during dataset acquisition, we applied a series of data augmentation techniques, including image scaling, shifting, and mirroring, to all training images. Our DICL model, which is based on the ICL architecture, was employed with the consideration that the original images were excessively high-resolution; hence, we randomly cropped them into 256×256 patches. Ultimately, the model exhibiting the highest evaluation metrics on the validation set was selected as our optimal model.

Table 2. Training protocols.

Network initialization	
Batch size	8
Patch size	1×256×256
Total epochs	1000
Optimizer	SGD
Initial learning rate (lr)	0.0001
Lr decay schedule	halved by 40 epochs
Training time	5.5 hours
Loss function	Dice Loss, MSE Loss, CE Loss

Table 3. Training protocols for second stage.

Network initialization	
Batch size	16
Patch size	1×256×256
Total epochs	100
Optimizer	SGD
Initial learning rate (lr)	0.0001
Lr decay schedule	halved by 40 epochs
Training time	2.5 hours
Loss function	Dice Loss, CE Loss

4 Results and discussion

4.1 Quantitative results on validation set

As shown in Table 4, we present the evaluation metrics calculated from online validation. Our model performs well on the validation set, with scores of 85.44% (image-level Dice), 74.99% (image-level IoU), 89.32% (image-level NSD), 23.85% (instance-level Dice), 55.14% (instance-level IoU), 66.53% (instance-level NSD) and 60.47% IA on the validation set. These metrics demonstrate superior performance compared to the classical segmentation algorithm U-Net [7] and the baseline model ICL that we used.

Furthermore, we conduct another set of ablation experiments to assess the impact of deformable convolution on model performance, as detailed in Table 4. When training with deformable convolution, we observe scores of 85.44% (image-level Dice), 74.99% (image-level IoU), 89.32% (image-level NSD), 23.85%

Table 4. Total quantitative evaluation results.

Method	image-level			instance-level			
	Dice (%)	IoU (%)	NSD (%)	Dice (%)	IoU (%)	NSD (%)	IA (%)
U-Net	80.52	68.55	84.39	15.71	49.78	60.55	53.60
ICL	81.51	69.84	85.43	16.90	45.93	56.06	48.52
DICL	85.44	74.99	89.32	23.85	55.14	66.53	60.47

(instance-level Dice), 55.14% (instance-level IoU), 66.53% (instance-level NSD) and 60.47% IA on the validation set. Without deformable convolution during training, the corresponding metrics are scores of 81.51% (image-level Dice), 69.84% (image-level IoU), 85.43% (image-level NSD), 16.90% (instance-level Dice), 45.93% (instance-level IoU), 56.06% (instance-level NSD) and 48.52% IA. This suggests that deformable convolution can more effectively concentrate on the characteristic information of teeth.

It is worth mentioning that our results are also validated on the online validation set, further confirming the excellent performance of our model.

4.2 Qualitative results on validation set

Good segmentation cases In case 0007 in Fig. 3, DICL demonstrates a more pronounced capability in semantic segmentation of the central teeth, accurately delineating different categories of central teeth, a task that ICL and U-Net fail to accomplish correctly. In case 0012, while both DICL and ICL can accurately categorize teeth, DICL offers a finer granularity in boundary segmentation, whereas U-Net struggles to even correctly identify the categories at this stage.

Failure case analysis In case 0006 and case 0020 in Fig. 3, the U-Net encountered difficulties in segmenting teeth with implants, and the two improved methods were unable to segment accurately. It can be explained that the presence of implants confuses all approaches.

4.3 Results on final testing set

You can describe this section like this, and we encourage you to add the necessary information: We obtained scores of 84.45% (image-level Dice), 73.60% (image-level IoU), 88.57% (image-level NSD), 22.36% (instance-level Dice), 57.39% (instance-level IoU), 69.85% (instance-level NSD) and 65.82% IA on the official test set. The average time latency and memory usage on the test set were 13.9 seconds with an average area under the GPU memory-time curve of 15088.5. Collectively, we ranked second among all submitted teams.

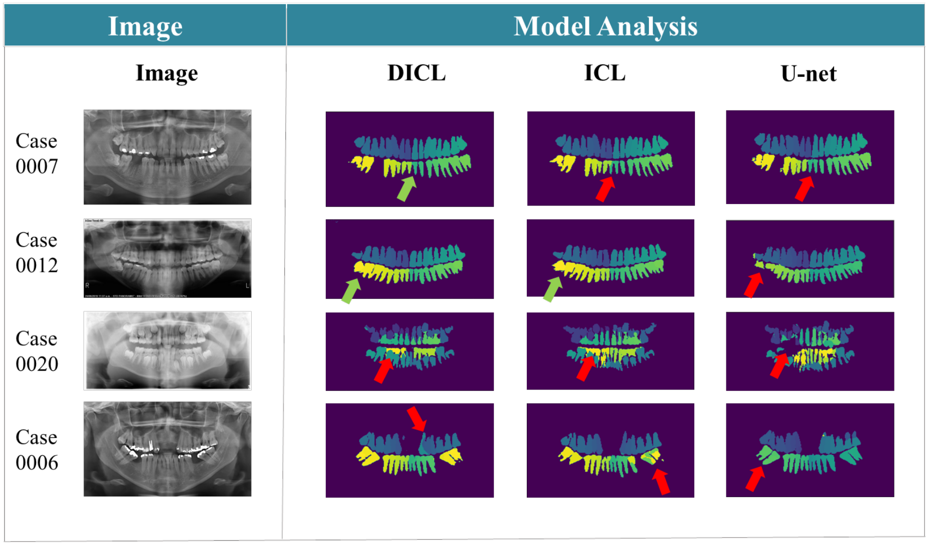

Fig. 3. Qualitative results on easy (case Validation 0007 and 0012) and hard (case Validation 0020 and 0006) examples.

4.4 Limitation and future work

In this study, we utilized the U-Net architecture as the backbone network framework for feature extraction and segmentation task learning. However, we have not delved into existing methods based on the transformer approach.

Looking ahead to future research, to further enhance the accuracy of our model, we plan to introduce methods based on transformer architectures or large model pre-training through in-depth experimentation. Additionally, it merits consideration to employ more advanced semi-supervised training models specifically designed for medical image segmentation tasks. Such refinements are anticipated to augment the model's representational capacity and generalization performance on complex medical images, thereby advancing the field.

5 Conclusion

In the current study, we employ the TB-FPN for tooth segmentation tasks, specifically applied to the dental panoramic radiographs dataset . Our algorithm performs remarkably well on the final test set, achieving a NSD score of 88.57% on the final test. Through thorough comparisons with algorithms such as U-Net and FPN, we observe that our algorithm demonstrated superior performance in terms of segmentation effectiveness.

Upon closer examination, we discover that our algorithm can accurately capture boundary regions in tooth segmentation tasks, particularly in accurately segmenting tooth roots, surpassing the accuracy of other comparative algorithms. Additionally, the TB-FPN enables segmentation with a finer granularity, allowing for more precise tooth segmentation. This is an aspect that other comparative algorithms fail to achieve.

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. Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E.: Semi-supervised and task-driven data augmentation. In: Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26. pp. 29–41. Springer (2019) 2
2. Dong, S., Luo, G., Tam, C., Wang, W., Wang, K., Cao, S., Chen, B., Zhang, H., Li, S.: Deep atlas network for efficient 3d left ventricle segmentation on echocardiography. Medical image analysis **61**, 101638 (2020) 2
3. Hao, Y., Liu, Y., Chen, Y., Han, L., Peng, J., Tang, S., Chen, G., Wu, Z., Chen, Z., Lai, B.: Eiseg: An efficient interactive segmentation annotation tool based on paddlepaddle. arXiv preprint arXiv:2210.08788 (2022) 1
4. He, Y., Yang, G., Chen, Y., Kong, Y., Wu, J., Tang, L., Zhu, X., Dillenseger, J.L., Shao, P., Zhang, S., et al.: Dpa-densebiasnet: Semi-supervised 3d fine renal artery segmentation with dense biased network and deep priori anatomy. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 139–147. Springer (2019) 2
5. Panfilov, E., Tiulpin, A., Klein, S., Nieminen, M.T., Saarakkala, S.: Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 2
6. Peng, J., Wang, Y.: Medical image segmentation with limited supervision: a review of deep network models. IEEE Access **9**, 36827–36851 (2021) 2
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 8
8. Wang, S., Cao, S., Wei, D., Wang, R., Ma, K., Wang, L., Meng, D., Zheng, Y.: Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9162–9171 (2020) 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) 11

10. Zhang, Y., Ye, F., Chen, L., Xu, F., Chen, X., Wu, H., Cao, M., Li, Y., Wang, Y., Huang, X.: Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. Scientific Data **10**(1), 380 (2023) 5

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)	5
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/No
A pipeline/network figure is provided	Fig. 2
Pre-processing	3
Strategies to use the partial label	3&4
Strategies to use the unlabeled images.	4
Strategies to improve model inference	5
Post-processing	5
The dataset and evaluation metric section are presented	6
Environment setting table is provided	Table 1
Training protocol table is provided	Table 2
Ablation study	9
Efficiency evaluation results are provided	Table 4
Visualized segmentation example is provided	Fig. 3
Limitation and future work are presented	Yes
Reference format is consistent.	Yes