SemiT-SAM: Building a Visual Foundation Model for tooth instance Segmentation on Panoramic Radiographs

Jing Hao[1[0000−0002−2305−1201]], Moyun Liu[2[0000−0002−4530−2606]], Lei He[2[0009−0007−7113−3707]], Lei Yao[3[0009−0007−0304−3056]], James Kit Hon Tsoi[1[0000−0002−0698−7155]], and Kuo Feng Hung[1[0000−0002−3971−3484]]

[1] Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China
[2] The Huazhong University of Science and Technology, Wuhan, China
[3] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China
hungkfg@hku.hk

Abstract. Automated tooth instance segmentation on dental radiographs is a crucial step in establishing digital dental workflows. However, unlike the realm of natural images, there is currently no visual foundation model that can implement tooth instance segmentation accurately. In this paper, we built the first visual foundation model, SemiT-SAM, for tooth instance segmentation. This foundation model was meticulously designed in terms of model architecture design, the training data corpus, and the semi-supervised learning strategy. The SemiT-SAM inherited the capability of the SAM and was trained on a large-scale dataset TSI15k via the label-guided teacher-student knowledge distillation strategy. Based on SemiT-SAM, we participated in the challenge of "MICCAI STS 2024: Panoramic X-ray Images", and achieved satisfying performance with scores of 90.52% (image-level NSD) and 86.89% (image-level Dice) on the validation set. The checkpoint and code of SemiT-SAM, as well as the training dataset TSI15k, are available at: https://github.com/isbrycee/SemiTNet

Keywords: Teeth Segmentation · Visual Foundation Model · Semi-supervised learning.

1 Introduction

Automated tooth instance segmentation in dental radiographs is crucial for developing digital workflows that support diagnosis and treatment planning across different dental specialties[9,7,8]. However, the large-scale data with pixel-level annotations for teeth segmentation task, which is crucial for training neural networks, is scarcely available. The annotation process is labor-intensive and time-consuming, and necessitates dentists' clinical expertise and knowledge. Hence, in the field of dental imaging, unlike the realm of natural images, there is currently

no visual foundation model that can implement tooth instance segmentation accurately.

In the field of dentistry, there is an abundance of unlabeled panoramic images available for training purposes. Some efforts have concentrated on efficient training strategies, such as semi-supervised learning. [21] introduced a semi-supervised learning framework for dental panoramic caries segmentation, dividing suspected areas with uncertainties into unlabeled data for reducing dataset labeling costs. [16] presented an efficient semi-supervised method for caries detection and segmentation that leverages a small set of labelled images for training the teacher model and a large collection of unlabelled images for training the student model. [4] proposed the T-Mamba, which proved that the simple pseudo-labeling strategy for unlabeled data can improve the accuracy and robustness of the model. The semi-supervised learning approach, which is an efficient training method that reduces the reliance on costly annotated data, has demonstrated its significant advantage in data-scarce scenarios.

Different from these methods that focus on exploring semi-supervised learning on teeth segmentation task, we aim to build a visual foundation model for tooth instance segmentation in terms of three aspects, including model architecture design, the training data corpus, and the semi-supervised learning strategy. Firstly, we employed the GEM [5], which adopted a query-based encoder-decoder architecture and inherited the capability of the SAM [11], as the basic neural architecture for the visual foundation model on tooth instance segmentation task due to its accurate performance and efficient inference ability. Secondly, a large-scale dataset, TSI15k [6], including a total of 16,317 panoramic radiographs (1589 labeled and 14,728 unlabeled images), was used for training our teeth segmentation foundation model. To the best of our knowledge, this is the largest public dataset for tooth instance segmentation to date. Lastly, we adopt a label-guided teacher-student knowledge distillation strategy proposed in [1] to effectively leverage unlabeled images as additional training signals and enhance the model's performance. The trained-well visual foundation model for tooth instance segmentation is termed as SemiT-SAM. This model, along with its code and the training dataset TSI15k, are available at: https://github.com/isbrycee/SemiTNet.

We participated in the challenge of "MICCAI STS 2024: Panoramic X-ray Images" using our teeth segmentation foundation model SemiT-SAM, and it achieved satisfying performance. Firstly, we fine-tuned our SemiT-SAM using 30 labeled data provided in this challenge. This step was necessary to align the category space because we had originally defined 32 classes in our teeth segmentation foundation model, whereas this challenge defined 52 classes. Afterwards, the unlabeled data were used for the purpose of semi-supervised training. The training protocols are the same as the way we trained our teeth segmentation foundation model. Finally, our SemiT-SAM obtained scores of 86.89% (image-level Dice), 77.63% (image-level IoU), 90.52% (image-level NSD), 84.93% (instance-level Dice), 67.59% (instance-level IoU), 76.85% (instance-level NSD) and 76.00% IA in the validation set.

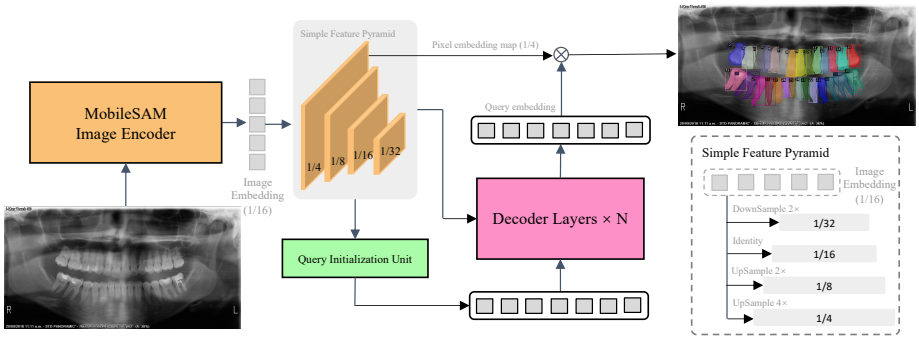


Fig. 1. The architecture of the foundation model SemiT-SAM features a streamlined encoder-decoder structure that includes four main components: an image encoder, a basic feature pyramid, a query initialization unit, and a mask decoder.

To summarize, our contributions are threefold:

- We proposed the first visual foundation model, SemiT-SAM, for tooth instance segmentation. This foundation model was meticulously designed from three aspects, including model architecture design, the training data corpus, and the semi-supervised learning strategy.
- The checkpoint and code of SemiT-SAM, as well as the dataset TSI15k, are available at: https://github.com/isbrycee/SemiTNet. We hope that these could be assets to further propel the application of artificial intelligence in the field of dentistry.
- Based on teeth segmentation foundation model SemiT-SAM, we achieved satisfying results on the validation set with scores 90.52% (image-level NSD) and 86.89% (image-level Dice).

2 Method

2.1 Preprocessing

To maintain the simplicity of the image preprocessing step, all panoramic radiographs are resized while preserving their aspect ratios, with the longer side adjusted to 1024 pixels. The shorter side is then padded with zeros to achieve a final image size of 1024 × 1024 pixels for network input.

2.2 Proposed Method

Model architecture The network GEM (Glass surface sEgMentor), previously proposed by our team for the glass surface segmentation, was selected as the basic neural architecture for the visual foundation model on tooth instance segmentation task due to its accurate performance and efficient inference ability. The SAM's backbone was utilized for extracting image features, and these

high-level image features that are beneficial to segmentation tasks are fed to an image decoder for obtaining instance masks and their corresponding categories. GEM adopts a query-based encoder-decoder architecture that consists of an image encoder, a simple feature pyramid, a query initialization unit, and a mask decoder, as shown in Fig. 1.

Considering the trade-off between the capability acquired from massive corpora inherited in the SAM and the efficient running time, we employ the image encoder ViT-Tiny [3] derived from MobileSAM [22]. The GEM uses only the last feature map from the image encoder to produce multi-scale feature maps via a simple feature pyramid following ViTDet [14]. Specifically, we generate feature maps of scales 1/8, 1/4, and 1/32 using deconvolution of strides 2 and 4 and maxpooling of strides 2, respectively. Given these hierarchical feature maps, we simply predict masks via a mask decoder used in MaskDINO [13] with simplified improvement. The query in the decoder was first initialized using a query initialization unit. Afterwards, the fusion operation for generating the pixel embedding map in MaskDINO is removed, and the feature map of scale 1/4 is directly appointed as the role of the pixel embedding map. Eventually, we obtain an output mask by dot-producting each content query embedding from the mask decoder with the pixel embedding map.

To summarize, an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ is inputted to the image encoder, and we can obtain four-scale feature maps *C2*, *C3*, *C4*, and *C5* via a simple feature pyramid $\mathcal{P}$, of which the resolutions are 1/4, 1/8, 1/16, and 1/32 of the input image, respectively. Then the mask decoder takes the queries $\mathcal{Q} \in \mathbb{R}^{N\times 256}$ and the flattened three high-level feature maps *C3*, *C4*, and *C5* as inputs and updates queries $\mathcal{Q}$. Finally, the updated queries $\mathcal{Q}$ dot-product with the pixel embedding map *C2* to obtain a predicted mask $\mathcal{M}$. The whole process can be formulated as follows:

$$C2,\ C3,\ C4,\ C5 = \mathcal{P}(\mathcal{E}(\mathcal{I})), \tag{1}$$

$$M = C2 \otimes \mathcal{D}(\mathcal{Q}, Flatten(C3,C4,C5)), \tag{2}$$

where $\mathcal{E}$ is the image encoder and $\mathcal{D}$ is the mask decoder. The $\otimes$ indicates the dot production [17]. Note that the prediction masks are output at each decoder layer.

Training dataset We carefully collected a large-scale dataset, TSI15k, from several open-sourced datasets for training the teeth segmentation foundation model. This dataset had been released in our previous work [6], and it comprises 1598 labeled data and 14k unlabeled data [4]. Duplicate images from the same institution across different datasets were excluded.

Semi-Supervised Training Semi-supervised learning can enhance model performance in situations with limited labeled data by leveraging unlabeled data

[4] This dataset is publicly available at https://huggingface.co/datasets/Bryceee/TISI15k-Dataset

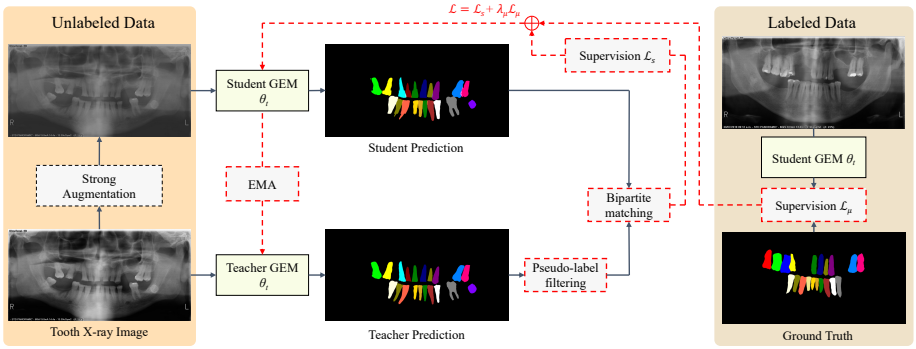


Fig. 2. Workflow of the distillation stage in the semi-supervised learning strategy. The original unlabeled panoramic radiographs were fed into the teacher model, while the strongly augmented unlabeled images were fed into the student model. The student model was updated using both the supervised loss (L_s) and unsupervised loss (L_u). The teacher model was subsequently updated using Exponential Moving Average (EMA).

[10,12,19,20]. The label-guided teacher–student knowledge distillation strategy [1] was employed to effectively leverage a large amount of unlabeled panoramic radiographs and build a stronger visual foundation model on tooth instance segmentation, which can be divided into three steps:

1. Teacher pre-training: The teacher model, parameterized by θ_t, is exclusively trained on annotated data.
2. Enhanced burn-in process: The student model, parameterized by θ_s, is initialized by the image encoder of MobileSAM [22] and trained on both labeled and unlabeled data using pseudo-labels generated by the teacher model in the first pre-training stage. During this phase, the teacher model is frozen.
3. In this stage, the weights of the student model are transferred to the teacher model, while the student continues to be trained on both labeled and unlabeled data as before. The teacher model is updated using an exponential moving average (EMA) [2] of the student's weights. The workflow for this stage is illustrated in Fig. 2.

The high-quality pseudo-label was generated using a simple thresholding method that takes into account both the predicted class probability and the size of the prediction. A predicted mask is chosen as a pseudo-label if it meets two criteria: (i) the maximum class probability is above the class threshold $p_c \geq \alpha_c$, and (ii) the size of the predicted mask is above the size threshold $\sum_{p=0}^{P_{H\times W}} \sigma(\hat{y}(p)) \geq \alpha_s$ where σ represents the sigmoid activation of the binary mask prediction. H and W refer to the height and width of the image, respectively. In our experimental settings, the class threshold α_c is 0.7 and the size threshold α_s is 5.

Loss function During the training phase, the total loss consists of the supervised and unsupervised losses, which share the same loss function, defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_{unsup}\mathcal{L}_{unsup} \tag{3}$$

The unsupervised loss weight λ_{unsup} is 2 in our experiments. The loss function is structured as a weighted sum of five loss components [15,24,18]:

$$\mathcal{L}_{sup/unsup} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{giou}\mathcal{L}_{giou} + \lambda_{focal}\mathcal{L}_{focal} + \lambda_{ce}\mathcal{L}_{ce} + \lambda_{Dice}\mathcal{L}_{Dice} \tag{4}$$

2.3 Post-processing

Due to the unique nature of tooth instance segmentation task, where no two identical categories of tooth instances exist within a single image, we have designed post-processing rules to ensure the accuracy of the segmentation results. Specifically, predicted instances with a confidence score below 0.45 are filtered out. Secondly, for the retained predictions, we ensure that only the mask with the highest confidence score is kept for each category.

3 Experiments

3.1 Dataset

The challenge of "MICCAI STS 2024: Panoramic X-ray Images" provides 2380 panoramic X-ray images in the training dataset and 20 panoramic X-ray images in the validation dataset [23]. The training dataset includes 30 cases with labels and 2350 unlabeled cases. The number of labeled data is obviously limited, so it is crucial to explore the semi-supervised training strategy on this challenge.

3.2 Evaluation metrics

Four metrics are employed to assess the model's performance, including Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD), mean Intersection-over-Union (mIoU), and Identification Accuracy (IA). The challenge evaluation criteria are not limited to segmentation accuracy but also include runtime and GPU memory consumption, providing a comprehensive assessment of segmentation algorithms.

3.3 Implementation details

Environment settings The environment settings are shown in Table 1.

Training protocols Firstly, we fine-tuned our teeth segmentation foundation model using the provided 30 labeled data in this challenge. This step was necessary to align the category space because we had originally defined 32 classes in our teeth segmentation foundation model, whereas this challenge defined 52 classes. Afterwards, the unlabeled data were used for semi-supervised training, which is described in Sec. 2.2. The detailed training protocols are demonstrated in Table. 2

Table 1. Development environments and requirements.

System	Ubuntu 18.04.6 LTS
CPU	Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz
RAM	16×4 GB
GPU (number and type)	Four NVIDIA GeForce RTX 3060 12G
CUDA version	11.6
Programming language	Python 3.8.19
Deep learning framework	torch 1.11.0, torchvision 0.12.0
Specific dependencies	MultiScaleDeformableAttention & Detectron2
Code	https://github.com/isbrycee/SemiTNet

Table 2. Training protocols.

Network initialization	Teeth Segmentation Foundation Model
Batch size	4
input image size	1024×1024
Total iterations	30000
Optimizer	AdamW
Initial learning rate (lr)	1e-4
Lr decay schedule	StepLR (decreased by 0.1 after 25k and 28k iters)
Training time	14 hours
Number of model parameters	21.6 M
Number of flops	107.3 G
Number of queries	100

4 Results and discussion

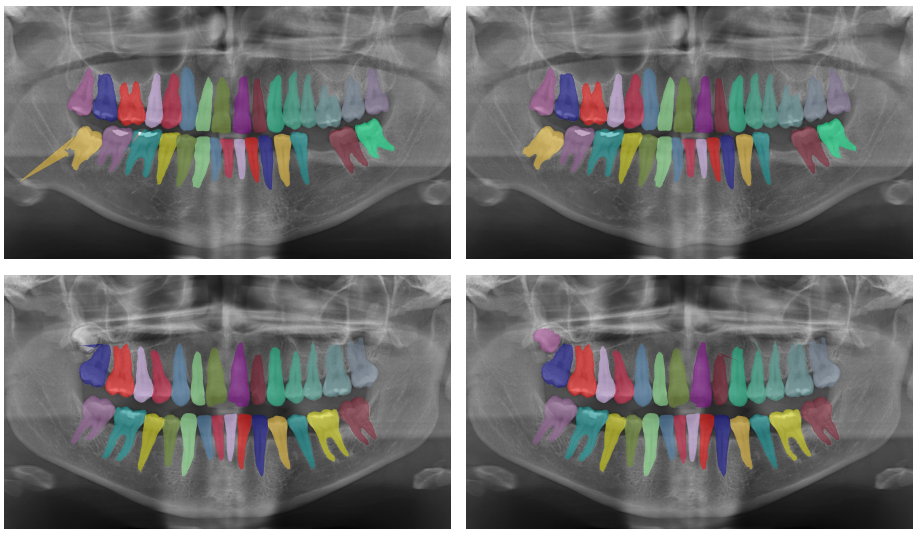
4.1 Quantitative results on validation set

The quantitative results are shown in Table 3. After fine-tuning our teeth segmentation foundation model with only 30 labeled panoramic radiographs, it achieved the following scores: obtained scores of 86.89% (image-level Dice), 77.63% (image-level IoU), 90.52% (image-level NSD), 84.93% (instance-level Dice), 67.59% (instance-level IoU), 76.85% (instance-level NSD) and 76.00% IA. This demonstrates that our teeth segmentation foundation model provides a robust prior for tooth instance segmentation capabilities. Based on the teeth segmentation foundation model, it is possible to align the category space using a relatively small number of labeled data and achieve highly accurate segmentation results.

Table 3. Quantitative evaluation results on validation set.

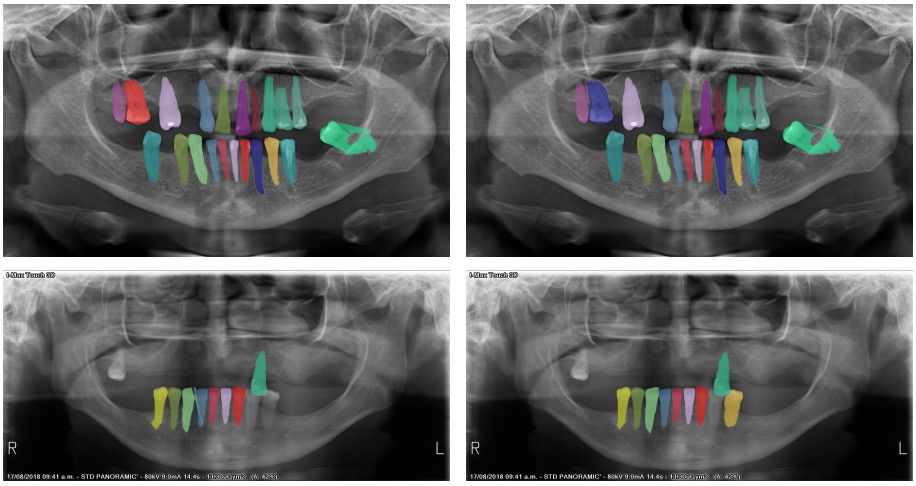
Method	image-level			instance-level			
	Dice (%)	IoU (%)	NSD (%)	Dice (%)	IoU (%)	NSD (%)	IA (%)
SemiT-SAM	86.89	77.63	90.52	**84.93**	67.59	76.85	76.00
SemiT-SAM$_{\text{unlabeled data}}$	**87.90**	**78.91**	**91.60**	83.88	**68.45**	**77.95**	**78.20**

With the help of unlabeled data and semi-supervised training strategy, the performance could be further improved. All metrics, with the exception of the instance-level Dice, have shown improvement, that is, the enhancements resulted in increases of 1.01% (image-level Dice), 1.28% (image-level IoU), 1.12% (image-level NSD), 0.86% (instance-level IoU), 1.10% (instance-level NSD), and 2.20% IA. We believe that since the teeth segmentation foundation model already provides a strong baseline for tooth instance segmentation capabilities, the improvement brought by unlabeled data is limited. Additionally, we suspect that due to the limited quality of pseudo-labels and the difficulty in adaptively controlling the threshold for generating pseudo-labels, the instance-level Dice metric, which is a pixel-level measure of the similarity between two masks, may decline.



(a) Before semi-supervised learning (b) After semi-supervised learning

Fig. 3. Two examples with successful segmentation results. (a) The predictive results of SemiT-SAM without using unlabeled data. (b) The predictive results of SemiT-SAM using unlabeled data via semi-supervised training strategy.



(a) Before semi-supervised learning (b) After semi-supervised learning

Fig. 4. Two failure examples. (a) The predictive results of SemiT-SAM without using unlabeled data. (b) The predictive results of SemiT-SAM using unlabeled data via semi-supervised training strategy. Although the use of unlabeled data can improve the accuracy of tooth segmentation to some extent, the segmentation performance remains poor in cases with a significant number of missing teeth.

4.2 Qualitative results on validation set

Good segmentation cases Figure 3 presents two examples that demonstrate satisfactory segmentation outcomes. When fine-tuning our teeth segmentation foundation model with only 30 labeled data samples, the results are relatively good but exist some bad cases, such as noticeable errors at the root boundaries (first row) and the inability to predict wisdom teeth (second row). After employing unlabeled data for semi-supervised training, these issues were resolved, thereby validating that unlabeled data can enhance model accuracy to a certain extent, particularly for the segmentation of detailed areas.

Failure case analysis Figure 4 presents two failure cases. In the cases of missing numerous teeth, the model struggles to accurately identify tooth numbering. Even after semi-supervised learning with unlabeled data, which leads to a marginal improvement in segmentation performance, there are still cases where some teeth remain unrecognizable.

4.3 Results on final testing set

We obtained scores of XX.XX% (image-level Dice), XX.XX% (image-level IoU), XX.XX% (image-level NSD), XX.XX% (instance-level Dice), XX.XX% (instance-level IoU), XX.XX% (instance-level NSD) and XX.XX% IA on the official test

set. The average time latency and memory usage on the test set were XX.XX seconds and XXXX MB, with an average area under the GPU memory-time curve of XXXX. Collectively, we ranked XXXX among all submitted teams.

4.4 Limitation and future work

When constructing the teeth segmentation foundation model, due to the fact that the training dataset TSI15k was entirely collected from publicly available datasets, the correctness of the annotations cannot be fully guaranteed. We believe that the performance of SemiT-SAM could be further enhanced if professional dental imaging experts could provide annotations for a set of high-quality tooth segmentation data.

In the future, we intend to further enhance the capabilities of SemiT-SAM and focus on the development of open-source oral multimodal large language models.

5 Conclusion

In this paper, we release the first visual foundation model, SemiT-SAM, for tooth instance segmentation. This foundation model was designed in terms of model architecture design, the training data corpus, and the semi-supervised learning strategy. Based on SemiT-SAM, we achieved satisfying results on the validation set with scores 90.52% (image-level NSD) and 86.89% (image-level Dice) in the challenge of "MICCAI STS 2024: Panoramic X-ray Images". Hoping our teeth segmentation foundation model SemiT-SAM could further propel the advancement of digital dentistry.

References

1. Berrada, T., Couprie, C., Alahari, K., Verbeek, J.: Guided distillation for semi-supervised instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 475–483 (2024) 2, 5
2. Cai, Z., Ravichandran, A., Maji, S., Fowlkes, C., Tu, Z., Soatto, S.: Exponential moving average normalization for self-supervised and semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 194–203 (2021) 5
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
4. Hao, J., He, L., Hung, K.F.: T-mamba: Frequency-enhanced gated long-range dependency for tooth 3d cbct segmentation. arXiv preprint arXiv:2404.01065 (2024) 2
5. Hao, J., Liu, M., Yang, J., Hung, K.F.: Gem: Boost simple network for glass surface segmentation via vision foundation models. arXiv e-prints pp. arXiv–2307 (2023) 2

6. Hao, J., Wong, L.M., Shan, Z., Ai, Q.Y.H., Shi, X., Tsoi, J.K.H., Hung, K.F.: A semi-supervised transformer-based deep learning framework for automated tooth segmentation and identification on panoramic radiographs. Diagnostics **14**(17), 1948 (2024) 2, 4
7. Hung, K.F., Ai, Q.Y.H., Wong, L.M., Yeung, A.W.K., Li, D.T.S., Leung, Y.Y.: Current applications of deep learning and radiomics on ct and cbct for maxillofacial diseases. Diagnostics **13**(1), 110 (2022) 1
8. Hung, K.F., Yeung, A.W.K., Bornstein, M.M., Schwendicke, F.: Personalized dental medicine, artificial intelligence, and their relevance for dentomaxillofacial imaging. Dentomaxillofacial Radiology **52**(1), 20220335 (2023) 1
9. Hung, K., Montalvao, C., Tanaka, R., Kawai, T., Bornstein, M.M.: The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. Dentomaxillofacial Radiology **49**(1), 20190107 (2020) 1
10. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016) 5
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023) 2
12. Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8300–8311 (2021) 5
13. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3041–3050 (2023) 4
14. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European conference on computer vision. pp. 280–296. Springer (2022) 4
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 6
16. Qayyum, A., Tahir, A., Butt, M.A., Luke, A., Abbas, H.T., Qadir, J., Arshad, K., Assaleh, K., Imran, M.A., Abbasi, Q.H.: Dental caries detection using a semi-supervised learning approach. Scientific Reports **13**(1), 749 (2023) 2
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) 4
18. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019) 6
19. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015) 5
20. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine learning **109**(2), 373–440 (2020) 5
21. Wang, X., Gao, S., Jiang, K., Zhang, H., Wang, L., Chen, F., Yu, J., Yang, F.: Multi-level uncertainty aware learning for semi-supervised dental panoramic caries segmentation. Neurocomputing **540**, 126208 (2023) 2

22. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023) 4, 5
23. 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) 6
24. Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y., Pan, Y.: Rethinking dice loss for medical image segmentation. In: 2020 IEEE International Conference on Data Mining (ICDM). pp. 851–860. IEEE (2020) 6

Table 4. 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
A pipeline/network figure is provided	Fig. 1 & Fig. 2
Pre-processing	3
Strategies to use the partial label	6
Strategies to use the unlabeled images.	5
Strategies to improve model inference	4
Post-processing	6
The dataset and evaluation metric section are presented	6
Environment setting table is provided	7
Training protocol table is provided	7
Ablation study	8
Efficiency evaluation results are provided	Table 3
Visualized segmentation example is provided	Fig. 3 & Fig. 4
Limitation and future work are presented	Yes
Reference format is consistent.	Yes