Self-Supervised Learning with Masked Image Modeling for Teeth Numbering, Detection of Dental Restorations, and Instance Segmentation in Dental Panoramic Radiographs

Amani Almalki Longin Jan Latecki
Department of Computer and Information Sciences, Temple University, Philadelphia, USA
{amani.almalki,latecki}@temple.edu

Abstract

The computer-assisted radiologic informative report is currently emerging in dental practice to facilitate dental care and reduce time consumption in manual panoramic radiographic interpretation. However, the amount of dental radiographs for training is very limited, particularly from the point of view of deep learning. This study aims to utilize recent self-supervised learning methods like SimMIM and UM-MAE to increase the model efficiency and understanding of the limited number of dental radiographs. We use the Swin Transformer for teeth numbering, detection of dental restorations, and instance segmentation tasks. To the best of our knowledge, this is the first study that applied self-supervised learning methods to Swin Transformer on dental panoramic radiographs. Our results show that the SimMIM method obtained the highest performance of 90.4% and 88.9% on detecting teeth and dental restorations and instance segmentation, respectively, increasing the average precision by 13.4 and 12.8 over the random initialization baseline. Moreover, we augment and correct the existing dataset of panoramic radiographs. The code and the dataset are available at https://github.com/AmaniHAlmalki/DentalMIM.

1. Introduction

The need for computer-assisted decisions is rising to facilitate diagnosis and treatment planning for dental care providers. Dental imaging is a valuable diagnostic tool for diagnosis and treatment plans, which is not possible solely through clinical exams and patient history [33]. A dental panoramic X-ray is a comprehensive tool that screens the teeth, surrounding alveolar bone and upper and lower jaws [28].

Moreover, dental restoration is a biocompatible synthetic material used to restore missing tooth structures. The missing tooth structure can be restored with full and partial coverage depending on the extension and intensity of the missing structure to restore the tooth's coronal (top) part. Furthermore, root canal filling is a restorative procedure used to fill the space inside the tooth structure (root portion) with biocompatible restorative materials. Various dental restorative materials are available in the dental world; each has its indication, advantages, disadvantages, and clinician preferences. Most dental restorative materials appear radiopaque in the x-ray, and they can be identified by dental care providers [1, 24].

However, manual intervention for teeth numbering and identification of tooth restorations is time-consuming and may overlook significant data. Thus, the interest in computer vision and computer science for automated processes was aroused. Few studies have attempted to apply computer vision algorithms in dental radiograph analysis. They include convolutional neural networks (CNNs) for teeth numbering and instance segmentation [21], two-stage network [36], Faster R-CNN [6, 15, 32, 30, 35], PANet [29], Mask R-CNN [14, 17, 18, 8], and U-Net network [27, 26, 16]. Recently, CNNs have enormous emerging applications in analyzing medical images with the advent of computation hardware/algorithm and expansion in the amount of data [21]. However, CNNs are limited in overall capability because of inherent inductive biases [11].

In this study, we propose to use a recently introduced Swin Transformer [22] to analyze dental panoramic radiographs. However, Swin Transformer requires large data for training, but there is only a very limited number of available dental radiographs. To alleviate this problem, we propose to use self-supervised learning. To the best of our knowledge, this is the first study that applied self-supervised learning methods to Swin Transformer on dental panoramic radiographs.

Recently, the self-supervised learning methods, SimMIM [34], UM-MAE [19], BEiT [2], MAE [13], SplitMask [12], MoCo v3 [9], and DINO [5], are effective in pretraining Transformers [11, 22] for learning visual represen-

tation. However, only UM-MAE and SimMIM pre-training methods are enabled for Pyramid-based ViTs with locality (Swin Transformer). Generally, the Masked Image Modeling (MIM) methods mask some image patches before they are fed into the transformer to predict the original patches in the masked area. This feature of aggregating information from the context helps many vision tasks. Although both UM-MAE and SimMIM provide a simple and efficient pre-training strategy for the Swin transformer encoder [22], the process of the input to the encoder is dissimilar. MAE discards the masked tokens and inputs only visible patches to the lightweight decoder. However, MAE also breaks the two-dimensional structure of the input image. Therefore, it is not applicable to the Swin transformer without the Uniform Masking (UM) introduced in [19] to bridge the gap between the MAE and Swin transformer. SimMIM includes the masked tokens in the encoder and uses them as a direct prediction mechanism. Using the randomly masked patches for SimMIM is a reasonable reconstruction target, and a lightweight prediction head is sufficient for pre-training. In addition, the location of the patches is essential in dental radiographs for a predictable outcome. SimMIM maintains the location of the patches known to both encoder and decoder, while MAE drops the location information, which may induce inaccuracy, as we demonstrate in this paper.

As there is no standard dental image dataset for pre-training (unlike ImageNet for natural images), SimMIM and UM-MAE are trained on the same dataset as the downstream tasks (excluding the test dataset). We conduct experiments on dental image tasks, including teeth numbering, detection of dental restorations, and instance segmentation on the dental panoramic X-rays dataset [29]. For these tasks, we use the base Swin Transformer (Swin-B) [22] as the backbone of Cascade Mask R-CNN [4]. We compare four Swin Transformer initializations, including SimMIM and UM-MAE, supervised initialization, and random initialization baseline. Our results show that SimMIM self pre-training can significantly improve object detection and instance segmentation performance on dental images.

Although previous studies have investigated teeth segmentation, we still address many gaps in this work. First, there is no comprehensive instance segmentation data set for teeth numbering. Previous work on the matter [29] used modified versions of binary semantic segmentation masks, which leads to a lack of instance overlapping and low-resolution outputs, resulting in inaccurate predictions, especially on the boundaries of the teeth. Second, there is a considerable amount of systematic errors because of the absence of dental expert supervision. Third, no prior work has simultaneously considered dental restoration segmentation besides tooth segmentation. The inclusion of teeth restorations increases the complexity of the computer vision problem because of class quantity and class imbalance.

To solve the data set issues, we augment and correct the existing dataset introduced in [29]. In addition to correcting the manual segmentation errors under expert supervision, we further expand the dataset by developing annotations for dental restorations, including direct restorations, indirect restorations, and root canal therapy. The labeling procedure resulted in a unique high quality, augmented dataset. Our data is available, upon request, under the name TNDRS (Teeth Numbering, Detection of Restorations, and Segmentation) annotations.

Our main contributions are twofold:

- We utilize self-supervised learning with SimMIM and UM-MAE to alleviate the problem of small data for panoramic radiographs.
- The corrected dataset leads to a significant increase in performance, while added labeling of dental restorations extends the horizon of possible dental applications.

2. Teeth numbering

In dentistry, various dental numbering systems are available for teeth numbering for adults and children. These numbering systems are universally accepted for better communication between dental care providers. The Universal Numbering System, Palmer Notation Numbering System, and Federation Dentiaure International numbering system (FDI) are the most commonly used system across the globe among dental professionals. The FDI system is the most widely used international system. In this system, every single tooth is assigned two-digit numbers; the first digit number represents each quadrant. The maxillary right and left quadrants are identified by the numbers 1 and 2, while the mandibular left and right quadrants are the numbers 3 and 4, respectively. The second digit numbers represent each tooth based on its location in the jaw from the middle. The central incisor is assigned to number 1, whereas the third molar is set to number 8 [32, 31].

3. Methods

The methods include two stages: the MIM pre-training and the downstream tasks, as illustrated in Fig. 1.

In the first stage, Swin Transformer is pre-trained with MIM self-supervised learning methods as the encoder. SimMIM divides the image into patches, replacing some random patches with mask tokens. Then, these patches, along with mask tokens, are input to the Swin encoder. Hence the positional encoding of both visible and masked patches is preserved, while UM-MAE drops those mask positions entirely. UM-MAE samples three random patches from each two-by-two grid, dropping 25% of the entire image.

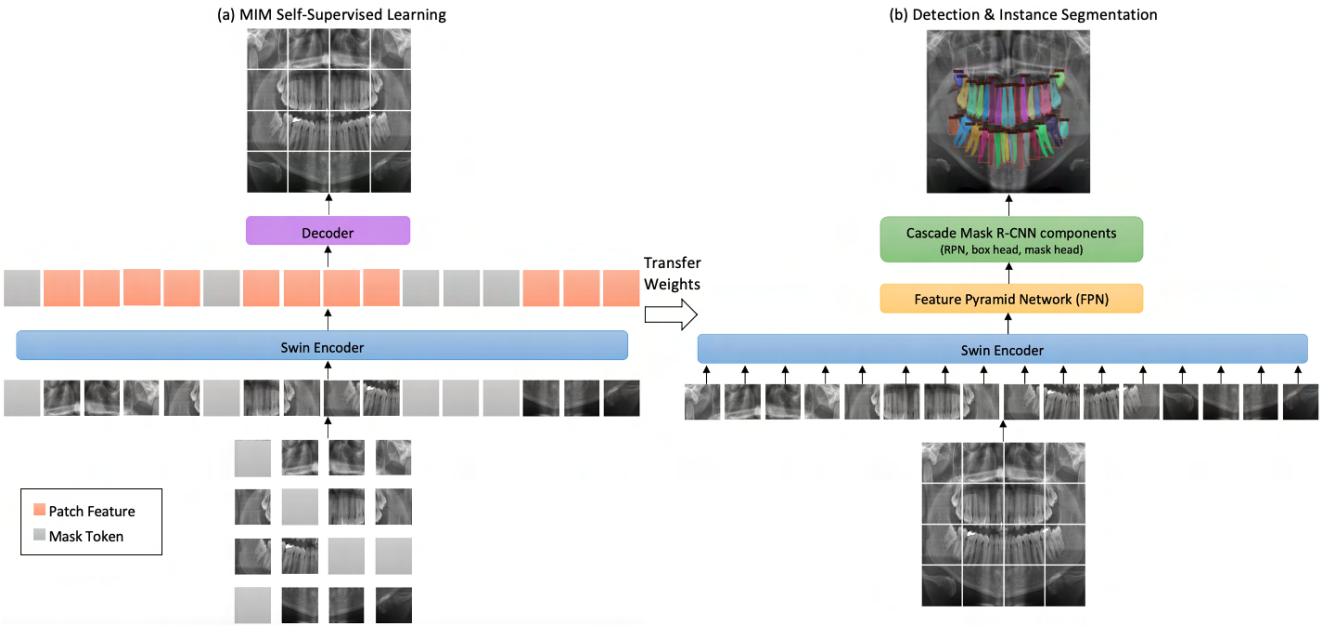

Figure 1. Pipeline for teeth detection, detection of dental restorations, and instance segmentation with MIM Self Pre-training. (a) A Swin Transformer is first pre-trained by MIM methods on the target dataset. (b) The pre-trained Swin Transformer is used as the backbone in Cascade Mask R-CNN with FPN for the detection and segmentation tasks.

Then it randomly masks 25% of the already sampled areas as shared learnable tokens. Finally, the sampled patches and the masked tokens are reorganized as a compact two-dimensional input under a quarter of the original image resolution to feed via the Swin encoder.

Then a decoder is appended to reconstruct the original patches at the masked area for both methods. In the second stage, the pre-trained Swin weights are transferred to initialize the detection and segmentation encoder. The features of the Swin Transformer backbone are fed to the neck (FPN [20]) and detection head (Cascade Mask R-CNN) for bounding box regression and classification as illustrated in Fig. 2. We select the Cascade Mask R-CNN [4] framework due to its ubiquitous presence in object detection and instance segmentation research. Then, the whole network is fine-tuned to perform the detection and segmentation tasks.

We use the base Swin Transformer backbone (Swin-B) and compare the effectiveness of four configurations as follows:

Random. The network is trained from scratch with randomly initialized weights, and no self-supervised methods are used. The Swin backbone configuration follows the code of [22], and the Cascade Mask R-CNN configuration uses the defaults in MMDetection [7].

Supervised. The Swin backbone is pre-trained for supervised object detection and instance segmentation using ImageNet-1K [10] images with their labels. We use the weights from [22] for Swin-B. Swin-B was pre-trained for 300 epochs.

SimMIM. We use the Swin-B weights pre-trained on self-supervised ImageNet-1K from [34]. This model was pre-trained for 100 epochs.

UM-MAE. Since ImageNet-1K pre-trained weights are not available; we use the official UM-MAE code release [19] to train Swin-B ourselves for 800 epochs (the default training length used in [19]) on unsupervised ImageNet-1K.

4. Experiments

4.1. Dataset augmentation and correction

TNDRS dental panoramic radiographs dataset. Detection, Numbering, and Segmentation (DNS) [29] is a dental panoramic X-rays dataset consisting of 543 annotated images with ground truth segmentation labels, including numbering information based on the FDI teeth numbering system. The image size is 1991x1127 pixels. The dataset annotations have some limitations as follows: 1) lack of instance overlapping; 2) some systematic errors because of the absence of dental expert supervision; 3) no segmentation of dental restorations. To overcome these issues, we modify and correct teeth instance segmentation and overlapping in all images. In addition, we contribute to further expanding the dataset by developing segmentation for dental restorations, including direct restorations, indirect restorations, and root canal therapy. This process was under a supervision of a dentist using the COCO-Annotator tool

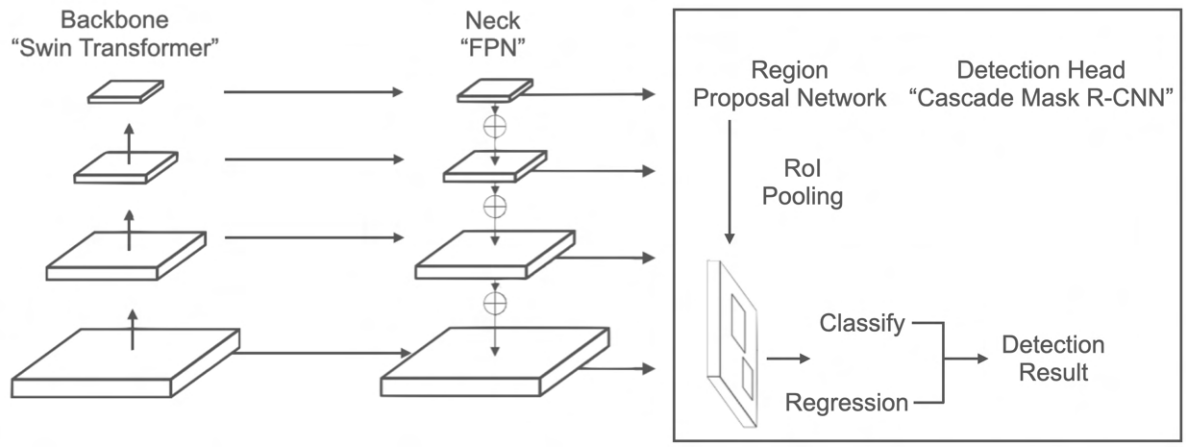

Figure 2. Illustration of the architecture for object detection.

[3]. We attended weekly meetings where related issues, such as numbering, dental restorations, and segmentation questions, were discussed. In the end, the annotations were reviewed to assure quality and avoid systematic and random errors. Fig. 3 shows a sample comparing the old and new versions of the dataset annotations, highlighting both the instance overlapping (blue arrow) and the correction of systematic errors (green arrow). Fig. 4 presents samples of segmentation of dental restorations.

We believe this is the most inclusive dataset for segmenting teeth and dental restorations in dental panoramic radiographs. We are providing our data, upon request, under the name TNDRS (Teeth Numbering, Detection of Restorations, and Segmentation) annotations.

4.2. Evaluation metric

For all our experiments, we split the data into five folds, each containing approximately 20% of the images. One of these folds is fixed as the test dataset (consisting of 111 images), and the other four folds (consisting of 108 images each) compose the training and validation datasets in a cross-validation manner. This process is repeated five times. The evaluation metric we adopt is the Average Precision for object detection and instance segmentation models.

4.3. Implementation details

Our experiments are implemented based on the PyTorch [25] framework and trained with NVIDIA Tesla Volta V100 GPUs. In all experiments, the batch size equals the total number of the training sample, which is 432. The input images are all resized to 800×600 pixels. We utilize the AdamW [23] optimizer in all experiments.

Data augmentation. We apply noise addition and horizontal flipping, which changes teeth numbers to their equivalent new values (left teeth numbers turned into the right numbers and vice-versa).

SimMIM pre-training. The base learning rate is set to 8e-4, weight decay is 0.05, $\beta 1$ = 0.9, $\beta 2$ = 0.999, with a cosine learning rate scheduler with warm-up for 10 epochs. We use a random MIM with a patch size of 16×16 and a mask ratio of 20%. We employ a linear prediction head with a target image size of 800×600 and use L1 loss to compute the loss for masked pixel prediction.

UM-MAE pre-training. The base learning rate is set to 1.5e-4, weight decay is 0.05, $\beta 1$ = 0.9, $\beta 2$ = 0.95, with a cosine decay learning rate scheduler with warm-up for 10 epochs. We use a random MIM with a patch size of 16×16 and a mask ratio of 25%. We employ a linear prediction head with a target image size of 800×600 and adopt mean squared error (MSE) to compute the loss for masked pixel prediction.

Task fine-tuning. For downstream tasks, we utilize single-scale training. The initial learning rate is 0.0001, and the weight decay is 0.05.

5. Results and analysis

SimMIM and UM-MAE reconstruction. The reconstruction results of SimMIM and UM-MAE are shown in Fig. 5. The five columns show the original images, the UM-MAE masked images, the UM-MAE reconstructed images, the SimMIM masked images, and the SimMIM reconstructed images. The results show that both MIM methods can restore lost information from the random context. It is worth noting that the ultimate goal of the MIM is to benefit the downstream tasks instead of generating high-quality reconstructions.

5.1. Quantitative results

Comparing initializations. Table 1 shows the results of teeth detection and instance segmentation only and com-

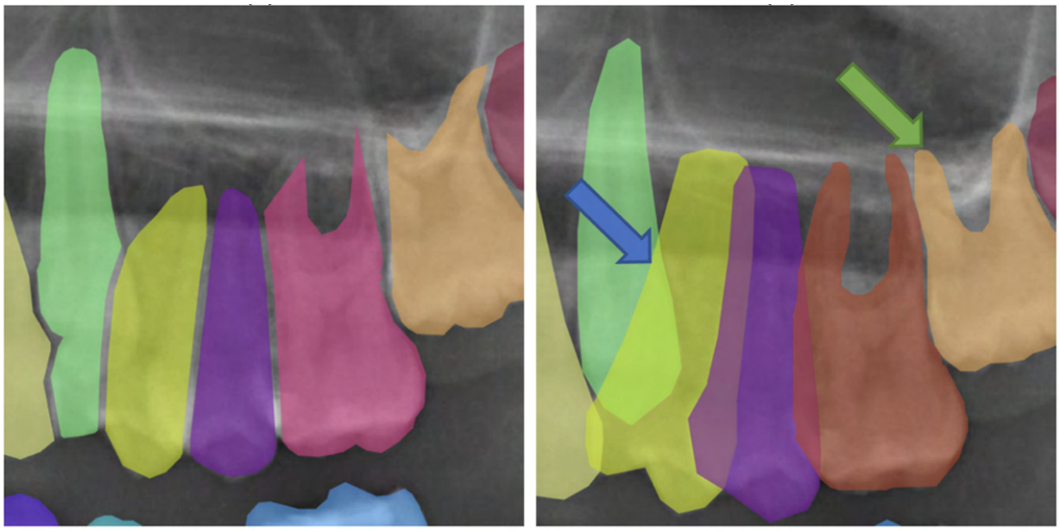

Figure 3. Comparison between the old and new dataset annotations. (a) Dataset old annotations. (b) Dataset new annotations. The blue arrow donates the inclusion of instance overlapping, while the green arrow indicates the correction of systematic errors, for example, unsegmented molar roots.

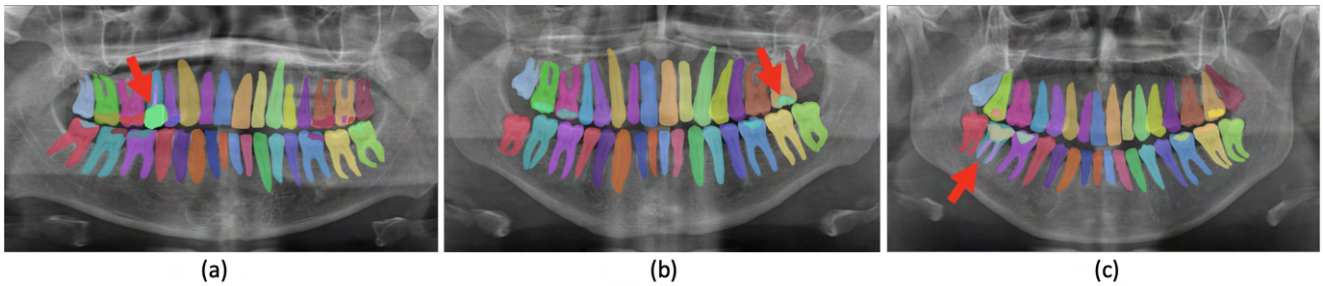

Figure 4. Samples of segmentation of dental restorations. Red arrows show an example of a) indirect restoration, b) direct restoration, and c) root canal therapy.

pares them to the previously published article from Silva *et al*. [29]. We present TNDRS fine-tuning results using the pre-trained models and random configurations described in Section 3. We make several observations.

(1) All four Swin Transformer initializations surpass the CNN-based SOTA of PANet with ResNet-50 backbone using ImageNet pre-training from Silva *et al*. [29].

(2) Fine-tuning from supervised IN-1K pre-training yields 3.4 higher AP^{box} than training from scratch (79.1 vs. 75.7) and 3.5 higher AP^{mask} (78.3 vs. 74.8).

(3) UM-MAE substantially outperforms supervised initialization by 5.4 AP^{box} (84.5 vs. 79.1), and 4.9 AP^{mask} (83.2 vs. 78.3).

(4) SimMIM outperforms UM-MAE by 1.6 AP^{box} (86.1 vs. 84.5), and 1.4 AP^{mask} (84.6 vs. 83.2).

Table 2 compares the four Swin Transformer initializations after data augmentation of dental restorations. Our results prove that the SimMIM method achieved the highest performance of 90.4% and 88.9% on detecting teeth and dental restorations and instance segmentation, respectively.

Parameter setting. In Table 3, we conduct experiments on teeth detection and instance segmentation tasks with different SimMIM pre-training epochs and mask ratios. First, the performance of SimMIM does not benefit from longer training. Second, unlike the high mask ratio [34] adopted in natural images, the downstream tasks show different preferences for the mask ratio. Both tasks are consistently improved with a decrease in mask ratio from 60% to 10%. The reason why this decrease facilitates the training may be attributed to the fact that the relevant features are small on panoramic X-rays.

Dataset correction. After we correct teeth segmentation on DNS discussed in Section 4.1, teeth detection and instance segmentation performance are remarkably improved by 5.9 AP^{box} and 6.4 AP^{mask} as shown in Table 4.

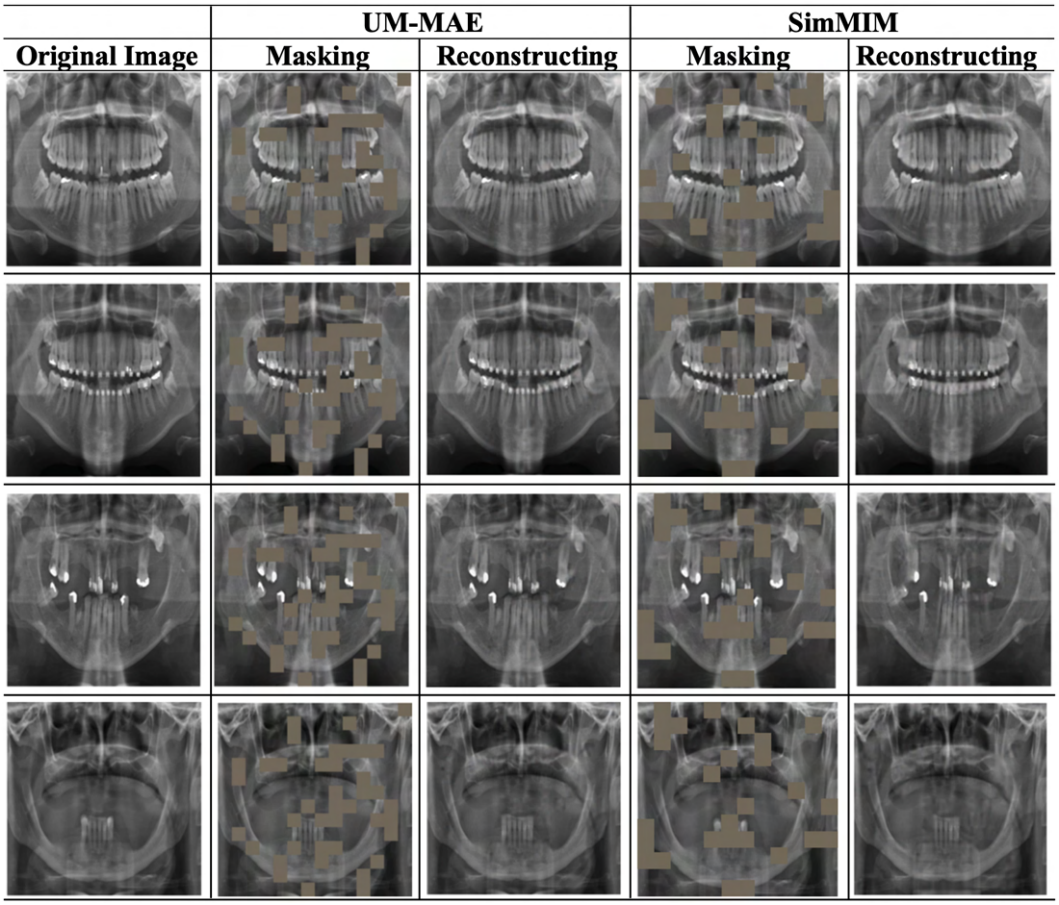

Figure 5. SimMIM and UM-MAE reconstruction results. The first column is the original image, and the second and fourth columns are the masked image where the masked region is denoted by gray patches. The third and fifth columns are the reconstruction of MIM from the unmasked patches.

Initialization	Backbone	Pre-training Data	AP^{box}	AP^{mask}
PANet[29]	ResNet-50	IN-1K w/ Labels	75.4	73.9
Random	Swin-B	None	75.7	74.8
Supervised	Swin-B	IN-1K w/ Labels	79.1	78.3
UM-MAE	Swin-B	IN-1K	84.5	83.2
SimMIM	Swin-B	IN-1K	**86.1**	**84.6**

Table 1. Results of teeth detection and instance segmentation only.

5.2. Qualitative results

In Fig. 6, the displayed results for four different images demonstrate qualitative samples of improved performance when Swin Transformer is pre-trained with SimMIM for teeth detection and segmentation only. These improvements in detection and segmentation agree with the quantitative results in Section 5.1.

Fig. 7 displays qualitative results after augmenting dental restorations when Swin Transformer is pre-trained with SimMIM.

5.3. Pre-training time and memory consumption

Comparing UM-MAE to the SimMIM framework, the core advantage of UM-MAE is the memory and runtime efficiency. In Table 5, we show their clear comparisons based on Swin-B. It is observed that UM-MAE speeds up by about 2× and reduces the memory by at least 2× against SimMIM, where their performances under the downstream tasks show the opposite.

Initialization	Backbone	Pre-training Data	AP^{box}	AP^{mask}
Random	Swin-B	None	77.0	76.1
Supervised	Swin-B	IN-1K w/ Labels	80.3	79.2
UM-MAE	Swin-B	IN-1K	88.3	85.7
SimMIM	Swin-B	IN-1K	**90.4**	**88.9**

Table 2. Results after augmenting dental restorations.

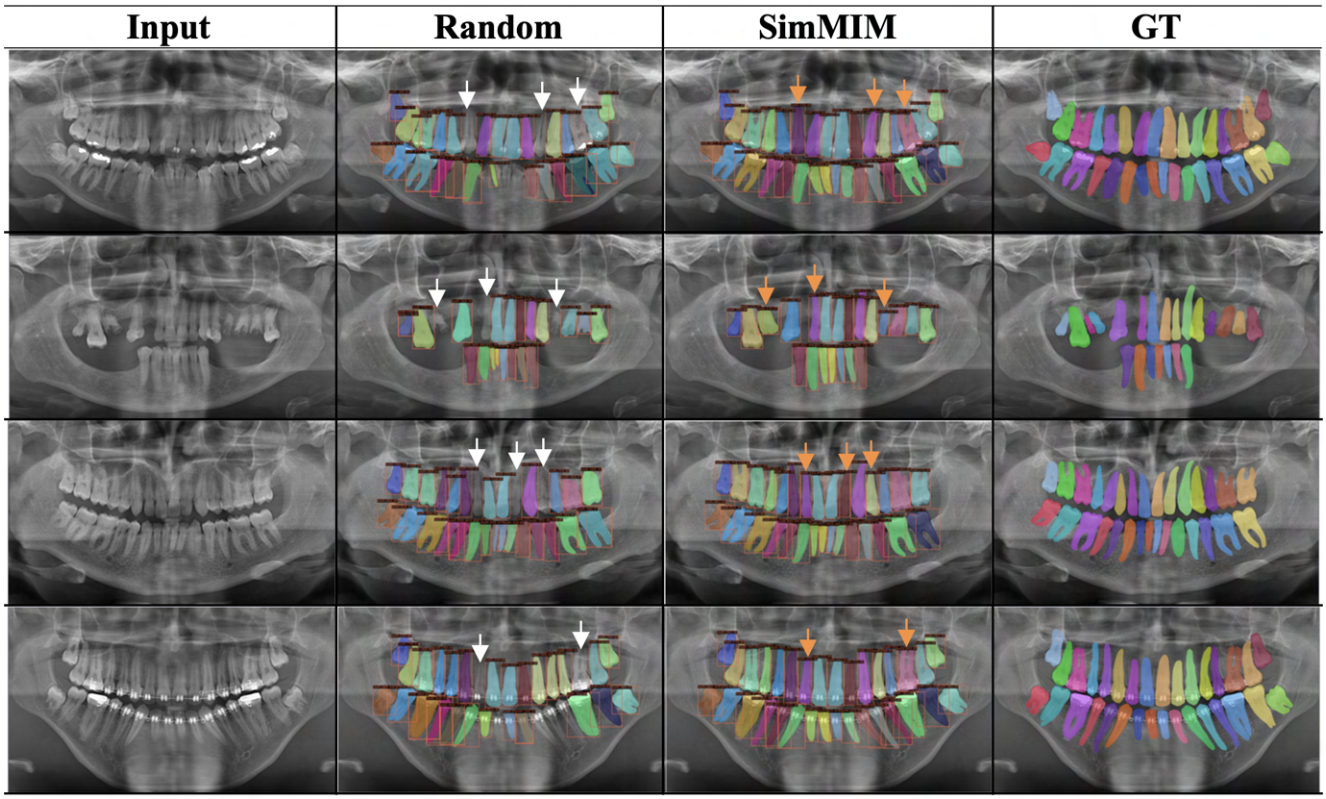

Figure 6. Qualitative results of teeth detection and instance segmentation only. Note that teeth detection and instance segmentation are missing (white arrows) when created by the baseline Swin Transformer approach compared to the segmentation produced by Swin Transformer pre-trained with SimMIM architecture (orange arrows).

Mask ratio	Pre-training Epochs	AP^{box}	AP^{mask}
60%	100	84.3	83.2
50%	100	84.7	83.6
50%	800	83.1	83.0
40%	100	85.5	83.9
30%	100	85.9	84.1
20%	100	**86.1**	**84.6**
10%	100	85.8	84.3

Table 3. The influence of Mask Ratios on teeth detection and instance segmentation tasks.

DNS Annotations	AP^{box}	AP^{mask}
Before Correction	80.2	78.2
After Correction	**86.1**	**84.6**

Table 4. Correction of teeth segmentation.

Method	Time	Memory
SimMIM	24.6 h	18.7 GB
UM-MAE	**12.5 h**	**6.7 GB**

Table 5. The comparison of pre-training time and memory consumption.

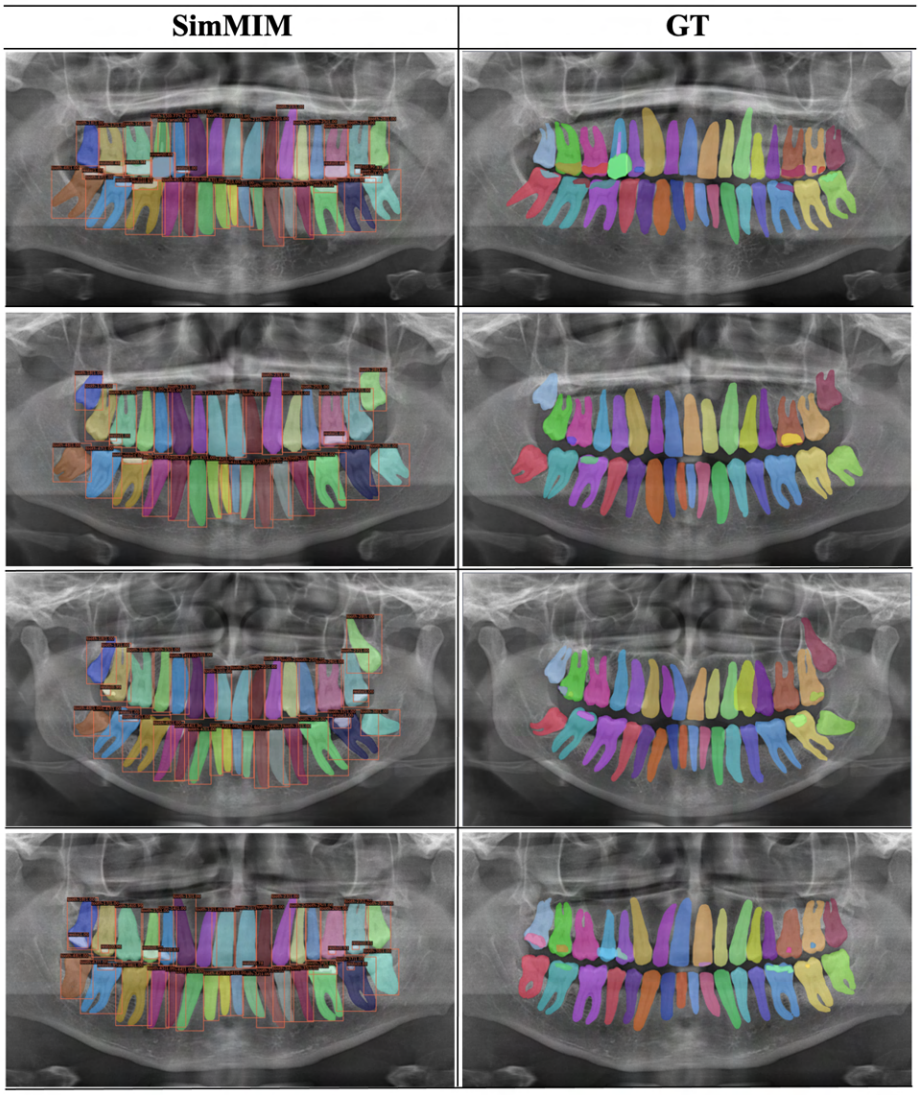

Figure 7. Qualitative results of detecting teeth and dental restorations and instance segmentation using SimMIM.

6. Conclusions

Two self-supervised learning methods were applied to Swin Transformer on dental panoramic radiographs: SimMIM and UM-MAE. The results of the masking-based method, SimMIM, obtained superior performance than UM-MAE, supervised and random initialization for detection of teeth, dental restorations, and instance segmentation. Based on this experiment, we can conclude that adjusting parameters, including mask ratio and pre-training epochs, is useful when applying SimMIM pre-training to the dental imaging domain for reliable outcomes. In addition, correcting the dataset annotations lead to further improvements that significantly surpass the available state-of-the-art results. Our plan for future work is to examine the efficacy of SimMIM pre-training in prognosis and outcome prediction tasks.

7. Acknowledgments

We would like to express our deepest thanks to Dr. Abdulrahman Almalki, a dental expert from the University of Pennsylvania, for his valuable discussions related to dentistry. This work was in part supported by the NSF Grant IIS-1814745. The calculations were carried out on HPC resources at Temple University supported in part by the NSF through grant number 1625061 and by the U.S. ARL under contract number W911NF-16-2-0189.

References

[1] Ragda Abdalla-Aslan, Talia Yeshua, Daniel Kabla, Isaac Leichter, and Chen Nadler. An artificial intelligence system using machine-learning for automatic detection and classification of dental restorations in panoramic radiography. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 130(5):593–602, 2020.

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Justin Brooks. COCO Annotator. `https://github.com/jsbroks/coco-annotator/`, 2019.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[6] Hu Chen, Kailai Zhang, Peijun Lyu, Hong Li, Ludan Zhang, Ji Wu, and Chin-Hui Lee. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Scientific reports*, 9(1):1–11, 2019.

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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.

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[14] Gil Jader, Jefferson Fontineli, Marco Ruiz, Kalyf Abdalla, Matheus Pithon, and Luciano Oliveira. Deep instance segmentation of teeth in panoramic x-ray images. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 400–407. IEEE, 2018.

[15] Changgyun Kim, Donghyun Kim, HoGul Jeong, Suk-Ja Yoon, and Sekyoung Youm. Automatic tooth detection and numbering using a combination of a cnn and heuristic algorithm. *Applied Sciences*, 10(16):5624, 2020.

[16] Thorbjørn Louring Koch, Mathias Perslev, Christian Igel, and Sami Sebastian Brandt. Accurate segmentation of dental panoramic radiographs with u-nets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 15–19. IEEE, 2019.

[17] Jeong-Hee Lee, Sang-Sun Han, Young Hyun Kim, Chena Lee, and Inhyeok Kim. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. *Oral surgery, oral medicine, oral pathology and oral radiology*, 129(6):635–642, 2020.

[18] André Ferreira Leite, Adriaan Van Gerven, Holger Willems, Thomas Beznik, Pierre Lahoud, Hugo Gaêta-Araujo, Myrthel Vranckx, and Reinhilde Jacobs. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. *Clinical oral investigations*, 25(4):2257–2267, 2021.

[19] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022.

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[21] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[24] Björn Molander. Panoramic radiography in dental diagnostics. *Swedish Dental journal. Supplement*, 119:1–26, 1996.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Dental x-ray image segmentation using a u-shaped deep convolutional network. In *International Symposium on Biomedical Imaging*, volume 1, pages 1–13, 2015.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] VE Rushton and K Horner. The use of panoramic radiology in dental practice. *Journal of dentistry*, 24(3):185–201, 1996.

[29] Bernardo Silva, Laís Pinheiro, Luciano Oliveira, and Matheus Pithon. A study on tooth segmentation and numbering using end-to-end deep neural networks. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 164–171. IEEE, 2020.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] DB Smith. The numbering of teeth. *New Zealand School Dental Service gazette*, 37(4):56, 1976.

[32] Dmitry V Tuzoff, Lyudmila N Tuzova, Michael M Bornstein, Alexey S Krasnov, Max A Kharchenko, Sergey I Nikolenko, Mikhail M Sveshnikov, and Georgiy B Bednenko. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. *Dentomaxillofacial Radiology*, 48(4):20180051, 2019.

[33] Stuart C White, Edward W Heslop, Lars G Hollender, Kristine M Mosier, Axel Ruprecht, Michael K Shrout, et al. Parameters of radiologic care: An official report of the american academy of oral and maxillofacial radiology. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 91(5):498–511, 2001.

[34] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[35] Kailai Zhang, Ji Wu, Hu Chen, and Peijun Lyu. An effective teeth recognition method using label tree with cascade network structure. *Computerized Medical Imaging and Graphics*, 68:61–70, 2018.

[36] Yue Zhao, Pengcheng Li, Chenqiang Gao, Yang Liu, Qiaoyi Chen, Feng Yang, and Deyu Meng. Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network. *Knowledge-Based Systems*, 206:106338, 2020.