IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, May 2024

MASKED DEEP EMBEDDINGS OF PATCHES FOR TEETH SEGMENTATION IN CONE BEAM CT IMAGES

Amani Almalki and Longin Jan Latecki

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

ABSTRACT

In modern dentistry, teeth localization and segmentation from dental cone-beam CT (CBCT) images are crucial for improving dental diagnostics, treatment planning, and population-based studies on oral health. However, creating automated algorithms for teeth analysis is a challenging task due to the limited availability of accessible data for training. This work extends the self-supervised learning framework of the masked autoencoder (MAE) transformer. While the MAE loss measures the quality of reconstructed masked patches, we propose to instead measure the closeness of the predicted deep embeddings of masked patches to their originals. This yields a better generalization ability on a very limited number of CBCT images, as documented by our results on teeth segmentation of CBCT images. We call our approach DEMAE for deep embedding MAE. Our results show that masking-based unsupervised learning methods may, for the first time, provide convincing transfer learning improvements on CBCT images, increasing the overall accuracy over both MAE and prior self-supervised pre-training.

***Index Terms*—** Self-supervised learning, Masked autoencoders, Tooth segmentation, Transformer, Deep embeddings

1. INTRODUCTION

In the last decade, digital dentistry has rapidly evolved, emphasizing the acquisition and division of complete three-dimensional (3D) tooth models. These models are crucial for defining the intended arrangement and movements of individual teeth, particularly for orthodontic diagnosis and treatment planning. Obtaining these comprehensive 3D tooth models presents a challenge. Currently, two main technologies for acquiring these models are intraoral or desktop scanning and cone beam computed tomography (CBCT) [1]. Intraoral or desktop scanning is convenient for capturing the surface geometry of tooth crowns but lacks information about tooth roots, essential for precise diagnoses and treatments. Conversely, CBCT provides comprehensive 3D volumetric data for all oral tissues, including teeth, and due to its high spatial resolution, it is widely used in oral surgery and digital orthodontics. This paper focuses on 3D tooth segmentation and identification from CBCT images, which crucial for digital orthodontics applications.

Segmenting teeth from CBCT images presents significant challenges due to several reasons. Firstly, in natural occlusion conditions where upper and lower teeth touch, it is difficult to differentiate and separate lower teeth from the opposing upper teeth along their occlusal surface due to a lack of variations in gray values [2, 3]. Similarly, distinguishing teeth from their surrounding alveolar bone is challenging due to their similar densities. Additionally, adjacent teeth with similar appearances pose confusion in identifying different teeth. Consequently, relying solely on the intensity variation of CT images, as attempted in previous tooth segmentation methods, has proven insufficient.

Prior attempts to address these issues involved using either the level-set method [2, 3, 4, 5] or template-based fitting methods [6] for tooth segmentation. The former methods necessitate a suitable initialization, often requiring laborious user annotations and yielding unsatisfactory results in natural occlusion conditions. The latter methods lack robustness when confronted with significant shape variations among different patients. While deep learning methods for medical image analysis [7, 8, 9] have shown promise in various tasks, their application to tooth segmentation has been limited.

Recent advancements in self-supervised learning have demonstrated the effectiveness of masked image modeling (MIM) [10, 11, 12] as a pre-training strategy for the Vision Transformer (ViT) [13] and the hierarchical Vision Transformer using shifted windows (Swin) [14, 15, 16]. MIM involves the masking and subsequent reconstruction of image patches, allowing the network to infer the masked regions by leveraging contextual information. We believe that the ability to aggregate contextual information is crucial in the context of CBCT image analysis. Among various MIM frameworks, the Masked Autoencoder (MAE) [11] stands out as a simple yet effective approach. MAE employs an encoder-decoder architecture, with a ViT encoder that receives only visible tokens and a lightweight decoder that reconstructs the masked patches using the encoder's patchwise output and trainable mask tokens.

We propose to use self pre-training since it is particularly advantageous in scenarios where acquiring suitable pre-training data is challenging. Additionally, self pre-training eliminates the domain discrepancy between the pre-training and fine-tuning stages by unifying the training data. Our experiments focus on teeth segmentation in 3D CT scans [17]. As our base model, we use UNEt TRansformer (UNTER) introduced in [18] for 3D CT scan analysis. Therefore, we call the proposed method UNETR+DEMAE. We apply UNETR+DEMAE pre-training on the same dataset that is used for the downstream task, i.e., to the training dataset.

Specifically, we propose to extend the self-supervised learning framework of the masked autoencoder (MAE) transformer [11]. While the MAE loss measures the quality of reconstructed masked patches, the loss of the proposed UNETR+DEMAE evaluates the predicted deep embeddings of masked patches. After pre-training, the decoder is discarded, and the encoder is applied to the downstream task, i.e., teeth segmentation. We compare three ViT Transformer initializations, including our proposed UNETR+DEMAE, MAE [11], and a transformer without any self-pre-training. The experimental results demonstrate that UNETR+DEMAE self-pre-training significantly enhances CBCT segmentation performance compared to the baselines. Our main contributions are threefold:

- We utilize self-supervised learning with masked autoencoders to alleviate the problem of small data for 3D CT scans.
- We replace the MAE reconstruction of masked patches with the reconstruction of patch embeddings. Hence, our loss is simply the L_2 distance between the predicted and computed embeddings over the masked patches.
- Our proposed method leads to a significant performance improvement. UNETR+DEMAE outperforms all state-of-the-art methods on the tooth segmentation task.

2. METHODS

2.1. Vision Transformer

Our framework utilizes the Vision Transformer (ViT) as the foundational architecture for both pre-training and subsequent tasks. The ViT comprises a patch embedding layer, position embedding, and Transformer blocks.

Patch Embedding: The patch embedding layer within the ViT is responsible for transforming data into sequences. Initially, 3D volumes $\mathbf{x} \in \mathbb{R}^{H\times W\times D\times C}$ are reshaped into a sequence of flattened 3D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^3\cdot C)}$. The parameters (H, W, D) represent the image resolution, (P, P, P) denotes the patch resolution, C signifies the input channel, and $N = HWD/P^3$ stands for the number of patches or the sequence length fed into the Transformer. These patches are then mapped to patch embeddings via a trainable linear projection.

Position Embedding: To retain positional information, the patch embeddings are supplemented with position embeddings. While the standard ViT utilizes 1D learnable position embeddings, our experiments led us to employ sine-cosine [11, 19] position embeddings during the pre-training stage. Sine-cosine functions provide a fixed pattern that is not learned during training. This can be advantageous to the model to learn more generalizable features and avoid overfitting to the specifics of the training data, which is very scarce. Subsequently, for downstream tasks, we initialize the learnable position embeddings with the sine-cosine embedding values.

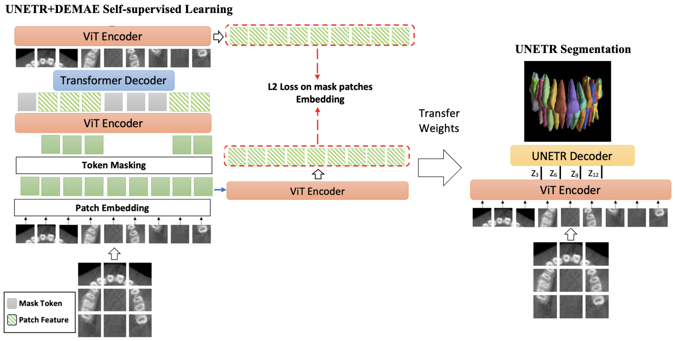

Fig. 1. **Segmentation Pipeline with MAE Self Pre-training.** Left: A ViT encoder is first pre-trained with MAE. A random subset of patches is input to the encoder and a transformer decoder reconstructs the full image. Right: The pre-trained ViT weights are transferred to initialize the segmentation encoder. Then the whole segmentation network, e.g., UNETR [18], is finetuned for segmentation.

Transformer Block: The ViT architecture involves layers comprising multiheaded self-attention (MSA) [20] and MLP blocks.

2.2. Self-Supervised Pre-training with Masked Autoencoders

This section delineates the constituents of the Masked Autoencoder (MAE): the encoder, the decoder, and the associated loss function.

Encoder. As illustrated in Fig. 1(Left), the ViT encoder is responsible for reconstructing the complete input data from partially masked patches. The input undergoes partitioning into non-overlapping patches, which are then randomly divided into visible and masked groups. The MAE encoder operates solely on visible patches, incorporating position embeddings to retain positional information. The resulting representation serves the purpose of reconstructing the masked input, urging the encoder to derive a comprehensive representation from partial observations.

Decoder. The MAE decoder is fed with a complete set of tokens, encompassing patch-wise representations from the encoder, alongside learnable mask tokens placed in the positions of masked patches. By integrating positional embeddings with all input tokens, the decoder aims to restore each specific patch within its masked position. It's noteworthy that the decoder serves as an auxiliary module exclusively for pre-training and is not utilized in downstream tasks.

Masked Sequence Generation. Patch embeddings are represented by a set E. Following the MAE approach, we randomly mask a subset of patches, represented as E_m, while unmasked embeddings are denoted as E_{um}. We replace the masked embeddings E_m with a shared learnable mask embedding E_{mask} without altering their positional embeddings. Finally, the corrupted embeddings E_c are formed by combining E_{um} with the sum of E_{mask} and a set of positional embeddings p. These corrupted embeddings are then inputted into the encoder for further processing.

Prediction. MAE [11] reconstructs the input by predicting the pixel values for each masked patch. Its loss function computes the mean squared error (MSE) between the reconstructed and original images in the pixel space. In contrast, We propose to compute the L_2 loss between the original and predicted embeddings of the mask patches. As our experimental results in Sec. 3.3 demonstrate, this leads to a performance increase.

2.3. Architectures for Downstream Tasks

Following MAE self-pre-training, we append task-specific head for the downstream task, i.e., tooth segmentation.

We employ the UNETR [18] built upon the pre-trained ViT encoder via MAE, in conjunction with a convolutional decoder initialized randomly. UNETR, designed for 3D image segmentation tasks, mirrors the concept of U-Net [21]. It involves skip connections between features from various encoder resolutions and the decoder. The input to the UNETR decoder constitutes a sequence of representations from the encoder. Each representation is reshaped to restore spatial dimensions, followed by iterative upsampling and concatenation with shallower features to enhance segmentation resolution.

3. EXPERIMENTS AND RESULTS

3.1. Datasets and Implementation Details

Tooth segmentation on CBCT images. We use the public dataset 3D CT scans [17]. There are a total of 150 CBCT images with a resolution varied from 0.25 mm to 0.35 mm. We randomly split the dataset into 80% for training and 20 for validation. The task is tooth segmentation from the 3D CT scans. Next, we normalize the intensity of the CBCT image to fit within the range of [0, 1]. For the creation of training data, we randomly extract 150 sub-volumes measuring $128 \times 128 \times 128$ around the alveolar bone ridge in the CT scan, resulting in approximately 18,000 sub-volumes for training. The dataset's ground truth includes annotations with tooth-level bounding boxes, masks, and labels.

During the testing phase, we employ the overlapped sliding window method to crop sub-volumes of size $128 \times 128 \times 128$ with a stride of $32 \times 32 \times 32$. Subsequently, in the scenario where two teeth segments overlap, we select the one with the highest value of $P_{cls} \times P_{id}$ as the final tooth prediction if the Intersection over Union (IoU) of their teeth segmentation results is greater than 0.2. Here, P_{cls} and P_{id} represent the probabilities for tooth classification and identification, respectively.

We conduct our experiments using PyTorch [22] and MONAI [23]. ViT-B/16 serves as the backbone, and we utilize AdamW as the optimizer across all experiments. The patch size for 3D volumes is set at $16 \times 16 \times 16$.

3.2. Evaluation metric

We use Dice similarity coefficient (DSC) to evaluate the performance of our model as follows:

$$DSC = \frac{2 \times |Y \cap Z|}{|Y| + |Z|}, \tag{1}$$

where Y and Z represent the voxelized predicted outcomes and the ground truth masks, respectively.

Additionally, we establish the accuracy of detection and identification as follows: assuming G represents the entirety of teeth within the ground truth data, and D indicates the set of teeth detected by our network, where within D there are L correctly labeled teeth. The detection accuracy (DA) and identification accuracy (FA) are determined through the following calculations:

$$DA = \frac{|D|}{|D \cup G|} \text{ and } FA = \frac{|L|}{|D \cup G|} \tag{2}$$

UNETR+DEMAE Self Pre-training. The starting learning rate (lr) remains at 1.5e-3, and the weight decay is set at 0.05. The learning rate decays to zero using a cosine schedule that includes warm-up periods. The pre-training for UNETR+DEMAE lasts for 100 epochs, utilizing training batch sizes of 256.

Finetuning for Teeth Segmentation. We apply a layer-wise learning rate decay (with a layer decay ratio of 0.75) to ensure the stability of UNETR training, along with implementing random DropPath with a 10% probability. The learning rate is set at 8e-3, and the training batch size is maintained at 256. Additionally, the learning rate during the fine-tuning phase also follows a cosine decay schedule.

3.3. Results

Teeth segmentation on CBCT images. Table 1 presents the quantitative results of tooth segmentation using various methods, and it clearly shows that UNETR+DEMAE outperforms other state-of-the-art methods.

Comparing the scores of DSC, PA and FA of UNETR with the other methods, it is evident that it achieves the highest performance, indicating its effectiveness in accurately segmenting tooth structures. This demonstrates the capability of the Transformers to capture relevant features and contextual information, leading to improved segmentation results.

The result of UNETR+MAE is superior to the standard UNETR, indicating further improvements. It also outperforms the ImageNet pre-training paradigm (UNETR+ImageNet). The combination of UNETR and MAE enhances the segmentation accuracy and ensures more precise delineation of tooth boundaries.

Our method, UNETR+DEMAE, surpasses the other methods and the standalone UNETR and its enhanced version MAE. Our method consistently achieves the highest results, highlighting the effectiveness of incorporating the loss on mask patch embeddings for tooth structure reconstruction.

Table 1. **Tooth Segmentation on CBCT scans.** UNETR+DEMAE self pre-training improves upon the UNETR baseline, ImageNet supervised pre-training, and MAE self-supervised learning.

Framework	DSC	DA	FA
U-Net(R50) [21]	84.18	82.84	79.19
AttnUNet(R50) [24]	85.92	63.91	79.20
TransUNet [25]	87.23	83.13	81.87
DSTUNet [26]	88.16	87.40	87.46
nnFormer [27]	86.07	80.17	86.57
nnUNet [28]	88.92	81.77	85.57
UNETR	89.46	90.88	88.03
UNETR+ImageNet	92.04	94.29	93.44
UNETR+MAE	93.01	95.25	94.37
UNETR+DEMAE	**94.20**	**99.65**	**97.57**

Parameter Setting. We perform experiments involving various UNETR+DEMAE pre-training epochs and mask ratios, as detailed in Table 2. Firstly, we note that the performance of UNETR+DEMAE does not improve with longer training periods. Secondly, unlike the high mask ratio commonly used in natural images [11], the segmentation task demonstrates varied preferences for different mask ratios. The most optimal segmentation outcomes are attained with a mask ratio of 25%.

Table 2. The influence of Mask Ratios and Pre-training Epochs on teeth segmentation of our UNETR+DEMAE.

Mask ratio	Pre-training Epochs	DSC	DA	FA
85%	100	91.32	96.43	95.59
75%	100	91.73	97.85	95.74
75%	800	90.14	97.32	94.89
50%	100	93.56	98.93	95.98
25%	100	**94.20**	**99.65**	**97.57**
10%	100	93.10	97.82	97.09

Qualitative results. Figure 2 presents qualitative examples that showcase the enhanced performance achieved on teeth segmentation through our UNETR+DEMAE pre-training in comparison to UNETR+MAE. The observed improvements in segmentation align with the quantitative findings shown in Table 1.

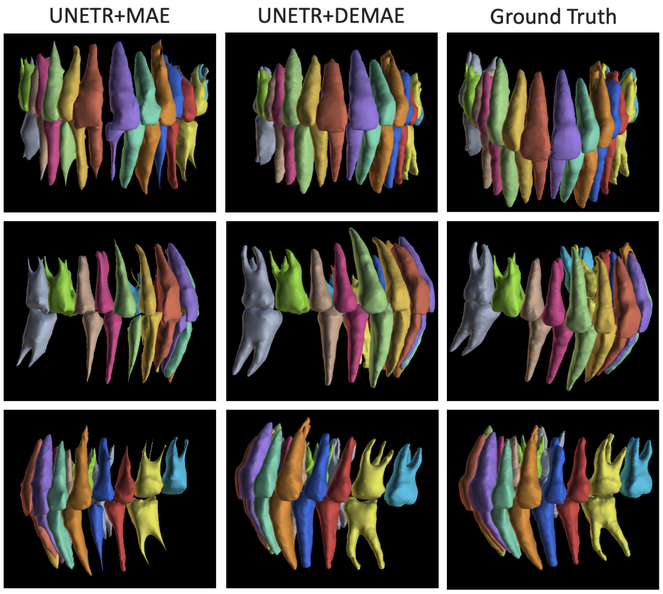

Fig. 2. Comparison of teeth segmentation of UNETR+DEMAE and baseline.

4. CONCLUSIONS

We have demonstrated that UNETR+DEMAE pre-training improves SOTA segmentation performance on 3D dental CT scan analysis. Importantly, UNETR+DEMAE self-pre-training outperforms existing methods on a small dataset, something that has not previously been explored. Our results also suggest that parameters, including mask ratio and strategy, should be tailored when applying masked autoencoders pre-training to the 3D dental scan domain. Together, these observations suggest that UNETR+DEMAE can further improve the already impressive performance of ViTs in CBCT scan analysis. In future work, we will test the efficacy of UNETR+DEMAE pretraining in prognosis and outcome prediction tasks.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [17]. Ethical approval was not required, as confirmed by the license attached to the open-access data.

6. ACKNOWLEDGMENTS

We would like to express our deepest thanks to Dr. Abdulrahman Almalki, a dental expert from the Department of Prosthetic Dental Science at Prince Sattam Bin Abdulaziz University, for his valuable discussions related to dentistry. Also, we would like to thank Dr. Zhiming Cui from Shanghai Tech University for sharing the dataset used in this study.

7. REFERENCES

[1] Lawrence Lechuga et al., "Cone beam ct vs. fan beam ct: a comparison of image quality and dose delivered between two differing ct imaging modalities," *Cureus*, vol. 8, no. 9, 2016.

[2] Mohammad Hosntalab et al., "Segmentation of teeth in ct volumetric dataset by panoramic projection and variational level set," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, pp. 257–265, 2008.

[3] Dong Xu Ji et al., "A level-set based approach for anterior teeth segmentation in cone beam computed tomography images," *Computers in biology and medicine*, vol. 50, pp. 116–128, 2014.

[4] Yangzhou Gan et al., "Tooth and alveolar bone segmentation from dental computed tomography images," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 196–204, 2017.

[5] Hui Gao et al., "Individual tooth segmentation from ct images using level set method with shape and intensity prior," *Pattern Recognition*, vol. 43, no. 7, pp. 2406–2417, 2010.

[6] Sandro Barone et al., "Ct segmentation of dental shapes by anatomy-driven reformation imaging and b-spline modelling," *International journal for numerical methods in biomedical engineering*, vol. 32, no. 6, pp. e02747, 2016.

[7] Qihang Yu et al., "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *CVPR*, 2018, pp. 8280–8289.

[8] Zizhao Zhang et al., "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *CVPR*, 2017, pp. 6428–6436.

[9] Zizhao Zhang et al., "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *CVPR*, 2018, pp. 9242–9251.

[10] Hangbo Bao et al., "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[11] Kaiming He et al., "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16000–16009.

[12] Zhenda Xie et al., "Simmim: A simple framework for masked image modeling," in *CVPR*, 2022.

[13] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[14] Ze Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10012–10022.

[15] Amani Almalki and Longin Jan Latecki, "Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs," in *WACV*, 2023, pp. 5594–5603.

[16] Amani Almalki and Longin Jan Latecki, "Enhanced masked image modeling for analysis of dental panoramic radiographs," in *ISBI*. IEEE, 2023.

[17] Zhiming Cui et al., "Hierarchical morphology-guided tooth instance segmentation from cbct images," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 150–162.

[18] Ali Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," in *WACV*, 2022, pp. 574–584.

[19] Xinlei Chen et al., "An empirical study of training self-supervised vision transformers," in *ICCV*, 2021, pp. 9640–9649.

[20] Ashish Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[22] Adam Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *NeurIPS*, 2019.

[23] MONAI Consortium, "MONAI: Medical Open Network for AI," 3 2020.

[24] Jo Schlemper et al., "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[25] Jieneng Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[26] Zhuotong Cai et al., "Dstunet: Unet with efficient dense swin transformer pathway for medical image segmentation," in *ISBI*, 2022.

[27] Hong-Yu Zhou et al., "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[28] Fabian Isensee et al., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, 2021.