One could ask the question of whether it is possible to simplify the architecture of DentalMAE proposed in Fig.1 by using the decoder to directly predict the embeddings of the masked triangle patch embeddings. Here we provide experimental evidence that such a simplification leads to performance degradation.

A. Simplified Architecture

The diagram of the simplified architecture, where the decoder is trained to directly predict masked triangle patch embedding, is shown in Fig. 5. We term this approach, DentalMAE-direct. Although both our DentalMAE and DentalMAE-direct take a complete set of tokens (including both encoded visible patches and mask tokens) as input to the decoder, the decoder's output trained as an embedding predictor shows lower performance, as indicated in Table 4.

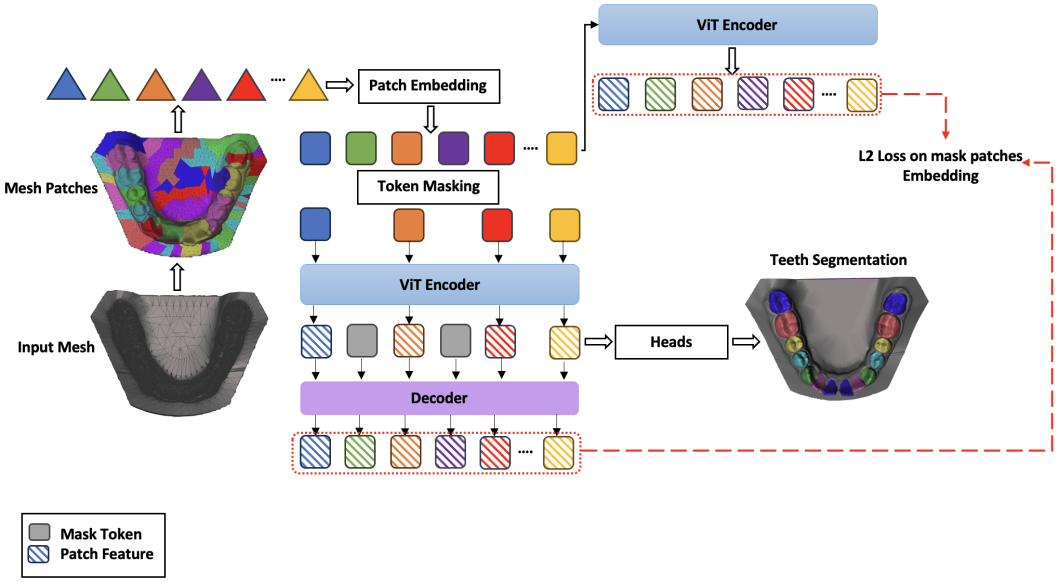

Figure 5. The teeth segmentation pipeline with DentalMAE-direct.

Method	OA	DSC	SEN	PPV
DentalMAE-direct	0.941	0.924	0.933	0.927
DentalMAE (ours)	**0.983**	**0.970**	**0.977**	**0.989**

Table 4. The decoder performance evaluation.

The discrepancy in performance between the DentalMAE and DentalMAE-direct could be attributed to several factors:

- In DentalMAE-direct the decoder is repurposed to a predictor, and the ViT encoder is missing, which reduces the number of parameters.
- In DentalMAE, the decoder first focuses on reconstructing the vertices and face features of the masked patches. This sequential reconstruction process allows the model to gradually incorporate contextual information from the visible patches, aiding in a better understanding of the masked patches. In DentalMAE-direct, however, the decoder directly predicts the patch embeddings without the benefit of the contextual information acquired during the reconstruction process.
- In DentalMAE, the decoder receives feedback and training signals from the reconstructed vertices and face features, which provide a more comprehensive supervision signal for the subsequent patch embedding prediction. This sequential reconstruction process offers more explicit guidance to the model during training, enabling it to learn better rep-

resentations. In DentalMAE-direct, the absence of this intermediate reconstruction step may result in a weaker or less informative training signal for the embedding prediction task, impacting the performance negatively.

- Without the sequential reconstruction process, DentalMAE-direct may face challenges in maintaining spatial coherence while predicting patch embeddings. The step-by-step reconstruction in DentalMAE helps the model capture the spatial relationships and coherence between the patches, leading to more accurate predictions. In DentalMAE-direct, where this reconstruction process is absent, the model may struggle to preserve spatial coherence, resulting in lower performance.