# STEAM: Self-supervised TEeth Analysis and Modeling for Point Cloud Segmentation

Yifan Liu[1], Chen Yang[2], Weihao Yu[1], Xinyu Liu[1], Hui Chen[3], Max Q.-H. Meng[4], Yixuan Yuan[1](✉)

[1] The Chinese University of Hong Kong
[2] The Hong Kong University of Science and Technology
[3] The University of Hong Kong
[4] Southern University of Science and Technology
yxyuan@ee.cuhk.edu.hk

**Abstract.** Accurate segmentation of 3D tooth point clouds from intraoral scanner (IOS) data is crucial for orthodontic applications. While current methods show promise, their reliance on high-quality labeled datasets is limited due to costly annotation processes, which further constrain their practical generalizability. We address this challenge with STEAM, a self-supervised learning framework that learns comprehensive features from large-scale unlabeled tooth point clouds. Built upon the masked autoencoder, our framework incorporates two key innovations: Gradient-guided Adaptive Masking (GAM), which adaptively identifies and prioritizes challenging regions by analyzing local feature variations during the training process, and Multi-attribute Geometric Reconstruction (MGR), which reconstructs multiple geometric attributes including point distributions, normals, and curvatures to capture geometric features of different granularity. Through extensive experiments on public datasets, our approach demonstrates superior performance in downstream segmentation tasks with minimal labeled data, achieving significant improvements over existing methods. The results validate STEAM effectiveness in maximizing the utility of limited labeled data for practical dental applications.

**Keywords:** Self-supervised Learning · 3D Tooth Segmentation · Point Cloud Analysis

## 1 Introduction

Recent advances in computer-aided-design (CAD) have established the Intra-Oral Scanner (IOS) [26] as a cornerstone technology in digital dentistry, enabling rapid 3D surface reconstruction for various clinical applications, including clear aligner design, dental restoration, and aesthetic smile enhancement. Precise tooth segmentation serves as a critical foundational step for these applications [12,17]. Due to the substantial point cloud data generated by IOS, typically comprising approximately 100,000 points per scan, manual segmentation

remains prohibitively resource-intensive [9]. This urgently requires automated, high-precision 3D tooth segmentation solutions in clinical practice.

Deep learning has revolutionized 3D tooth segmentation with numerous pioneering approaches [27,24,13,6,9,23,16], advancing both local feature extraction capabilities and specialized architectural designs. Despite these innovations, the reliance on limited annotated datasets constrains their clinical applicability, particularly in generalizing to diverse 3D scans. While expanding annotation coverage [9,23] offers partial improvement, the substantial costs and effort required for manual labeling remain prohibitive. Given the abundant availability of unlabeled 3D scans, self-supervised pre-training emerges as a promising solution to enhance model generalization efficiently. Self-supervised pre-training has demonstrated remarkable success across multiple domains, from natural language processing [7,21,3] to computer vision [4,11,10,1] and 3D understanding [28,18,25]. Notably, the masked autoencoder (MAE) paradigm [1,10,18] has proven particularly effective for large-scale applications. This approach strategically masks high proportions of input patches and leverages Transformer architectures [8] to reconstruct masked regions, enabling robust feature learning from unmasked contexts. The knowledge encoded during large-scale pre-training significantly enhances downstream task performance through fine-tuning, offering a powerful foundation for specialized applications.

Nevertheless, it is non-trivial to apply general MAE methods [25,18,14] directly on the tooth point cloud, caused by two primary challenges. First, dental scans predominantly consist of gingival points, causing random masking strategies to select gingival patches frequently. These patches, characterized by flat surfaces, provide minimal geometric information during reconstruction, limiting meaningful feature learning for encoders. This necessitates a selective masking strategy targeting geometrically complex regions, and also progressively adjusts the selection criteria along the learning process. Second, existing point cloud MAEs focus primarily on reconstructing basic spatial distributions of masked regions, i.e. point clouds. While this captures coarse morphological features, downstream tasks like dental segmentation require geometry attributes with different granularity for precise boundary delineation. Therefore, reconstruction targets incorporating diverse geometric properties need to be developed.

To tackle these challenges, we propose STEAM, a Self-supervised TEeth Analysis and Modeling framework for point cloud segmentation. Our framework introduces two innovative components: *Gradient-guided Adaptive Masking* (GAM) and *Multi-attribute Geometric Reconstruction* (MGR), designed to effectively mask challenging regions and reconstruct them with multiple geometric attributes. The main contributions can be summarized as follows: (1) We develop an adaptive masking mechanism where a teacher network assesses patch feature gradients to identify challenging regions, guiding the student network to focus on reconstructing these informative patches for more meaningful feature learning. (2) We design multiple specialized decoders to reconstruct diverse geometric attributes, including point distributions, surface normals, and curvatures, enabling the encoder to capture fine-grained surface characteristics at different

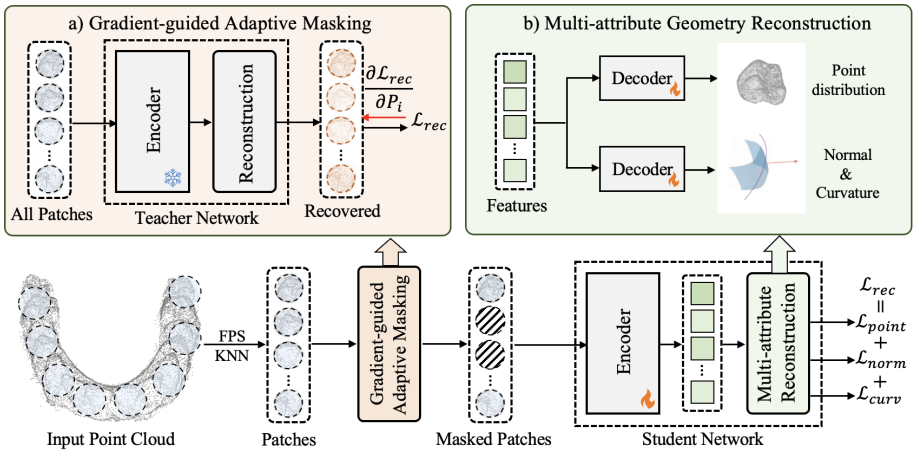


**Fig. 1.** Illustration of the proposed STEAM framework, including a) Gradient-guided Adaptive Masking and b) Multi-attribute Geometry Reconstruction.

granularities. (3) We validate our approach through extensive experiments on large-scale unlabeled dental scans for pre-training. Further fine-tuning experiments on public datasets demonstrate that STEAM significantly outperforms existing methods with minimal labeled data, providing a practical solution for clinical tooth segmentation applications.

## 2 Method

Our framework is designed for self-supervised learning for tooth point cloud analysis, by leveraging the scalable MAE (Sec. 2.1). As shown in Fig. 1, it begins with the Gradient-guided Adaptive Masking to represent the input point cloud as groups of patches and adaptively mask the informative patches based on the gradients of the teacher network (Sec. 2.2). Then, Multi-attribute Geometry Reconstruction is used to facilitate feature learning by reconstructing the masked geometry attributes with different granularity, including point distribution, normals, and curvatures (Sec. 2.3). After that, the pre-trained encoder is equipped with a segmentation decoder for the downstream tooth point cloud segmentation, optimized by the segmentation constraints (Sec. 2.4).

### 2.1 Preliminaries of Masked Modeling

MAE [1,10,18] aims to learn the latent features by reconstructing the masked signals. For the point cloud input, mainstream frameworks [25,18,14] typically utilize the Farthest Point Sampling (FPS) algorithm to sample patch centers, and search for corresponding nearest neighbors to compose patches. Then, a large ratio of the patches are randomly masked and the remaining unmasked ones

are tokenized by PointNet [19] and passed to a standard Transformer encoder [8], encoding high-level latent features. After that, these unmasked features are fed to a lightweight decoder, reconstructing the rough point distributions of the masked patches. This procedure enforces the encoder to encode meaningful shape representations, which can benefit the downstream tasks.

### 2.2 Gradient-guided Adaptive Masking

The random masking strategy used in existing MAEs [25,18,14] would mask a large portion of simple patches, caused by the huge number of gingiva points (around 40% of the tooth point cloud). As reconstructing such simple patches can not fully push the encoder to learn meaningful features, we propose to design a gradient-guided strategy to evaluate the reconstruction difficulty of each patch, and priorly reconstruct the patches with larger gradients. In doing so, the encoder is forced to extract informative features for reconstructing hard patches.

In particular, given input point cloud $P \in \mathcal{R}^{N\times 3}$ composed of $N$ points, we first use the Farthest Point Sampling (FPS) and K-Nearest Neighbors (KNN) algorithms to get grouped patches $G = \{g_1, g_2, ..., g_M\} \in \mathcal{R}^{M\times 3K}$, where $g_i \in \mathcal{R}^{3K}$ refers to the $i$-th patch containing $K$ points. Then, we do not mask patches and instead feed all patches to a frozen teacher network that shares parameters with the student network, obtaining the reconstruction loss $\mathcal{L}_{rec}$. After that, to evaluate the reconstruction difficulty of each patch $g_i$, we compute the gradients of $\mathcal{L}_{rec}$ back-propagated to each patch $g_i$, obtaining $\frac{\partial \mathcal{L}_{rec}}{\partial g_i}$. Considering patches that are harder to reconstruct would possess larger gradients, we select patches with the top-k gradients as the masked patches $G_m \in \mathcal{R}^{N_m\times 3K}$, and the remaining ones are treated as unmasked $G_u \in \mathcal{R}^{N_u\times 3K}$. The process is formulated as:

$$G_m = \{g_i | \frac{\partial \mathcal{L}_{rec}}{\partial g_i} \in TopK(\frac{\partial \mathcal{L}_{rec}}{\partial g_1}, \frac{\partial \mathcal{L}_{rec}}{\partial g_2}, ..., \frac{\partial \mathcal{L}_{rec}}{\partial g_M})\}, G_u = G/G_m. \quad (1)$$

It is worth noting that at the beginning of the training period, the teacher network shared from the student network can hardly reconstruct any patches, thus the patches are similarly difficult and the above masking strategy behaves more like random masking. With the training going on, the student network can learn latent features to reconstruct simple patches, the gradients derived from the teacher network would reflect the reconstruction difficulty reasonably, thus the masking strategy can mask harder patches, as we expected.

### 2.3 Multi-attribute Geometry Reconstruction

Existing MAE frameworks for point clouds [25,18,14] primarily focus on reconstructing point distribution alone, which limits the encoder to learning coarse geometric representations. However, in dental applications, accurate tooth identification requires the encoder to capture more fine-grained geometric characteristics, as different teeth often exhibit subtle variations in their surface properties

such as normals and curvatures. To address this limitation, we propose a Multi-attribute Geometric Reconstruction strategy that jointly reconstructs both point distribution and surface properties.

**Point Distribution Reconstruction** Following the patch generation process described in Sec. 2.2, we obtain unmasked patches $G_u \in \mathcal{R}^{N_u \times 3K}$ and masked patches $G_m$, where $N_u$ denotes the number of unmasked patches. The unmasked patches are processed through a standard Transformer encoder to obtain latent features $F_u \in \mathcal{R}^{N_u \times D}$, where $D$ represents the feature dimension. Subsequently, inspired by [10,18], we employ a lightweight decoder that takes the encoded features $F_u$ and randomly initialized masked tokens $T_m \in \mathcal{R}^{N_m \times D}$ as input, where $N_m$ represents the number of masked patches. The decoder generates the predicted point distribution $\hat{G}_m$ for the masked patches. The prediction is optimized using the chamfer distance loss:

$$\mathcal{L}_{point} = \sum_{x \in G_m} \underset{y \in \hat{G}_m}{min} ||x - y||_2^2 + \sum_{x \in \hat{G}_m} \underset{y \in G_m}{min} ||x - y||_2^2. \tag{2}$$

The reconstruction of point distribution enables the encoder to capture coarse geometric representations, which serve as essential features for the downstream segmentation task. Beyond point distribution, we further incorporate the reconstruction of surface properties below.

**Surface Properties Reconstruction** To provide more fine-grained surface constraints, we further add a lightweight decoder to reconstruct the surface properties of each patch, including normals $\hat{\mathcal{N}} = \{\hat{n}_1, \hat{n}_2, ..., \hat{n}_{N_m}\} \in \mathcal{R}^{N_m \times 3}$ and curvatures $\hat{\mathcal{C}} = \{\hat{c}_1, \hat{c}_2, ..., \hat{c}_{N_m}\} \in \mathcal{R}^{N_m}$ of the masked patch centers: $\hat{\mathcal{N}}, \hat{\mathcal{C}} = \mathit{Decoder}(F_u, T_m)$. To impose constraints on the reconstructed surface properties, we extract the normals $\mathcal{N}$ and curvatures $\mathcal{C}$ from the input IOS mesh using the algorithm in [5] as the ground truth, and design different losses for the normals and curvatures. For normals, we expect the predicted normal vectors to be aligned with ground truth, thus a cosine distance loss is used:

$$\mathcal{L}_{norm} = \frac{1}{N_m} \sum_{i=1}^{N_m} (1 - \frac{n_i \cdot \hat{n}_i}{||n_i||_2 ||\hat{n}_i||_2}). \tag{3}$$

For the curvature, we first use the Sigmoid function to re-scale the curvature range into [0, 1] for numeral stability, and then use the MSE loss to measure the difference:

$$\mathcal{L}_{curv} = \frac{1}{N_m} \sum_{i=1}^{N_m} ||\mathit{Sigmoid}(c_i) - \mathit{Sigmoid}(\hat{c}_i)||_2^2. \tag{4}$$

In summary, the total loss includes point distribution constraints, normal vector constraints, and curvature scalar constraints:

$$\mathcal{L}_{rec} = \lambda_1 \mathcal{L}_{point} + \lambda_2 \mathcal{L}_{norm} + \lambda_3 \mathcal{L}_{curv}, \tag{5}$$

where $\lambda_{1:3}$ are balancing weights. By optimizing the network with multiple targets, the encoder is encouraged to learn both the coarse point distribution and fine-grained surface properties, which can further benefit the downstream segmentation task.

### 2.4 Point Cloud Segmentation with Fine-tuning

The pre-trained encoder is further fine-tuned in a supervised manner for the downstream segmentation task. To obtain point-wise predictions, a randomly initialized decoder and a segmentation head are added after the pre-trained encoder for the feature propagation, as in previous works [25,18]. During the fine-tuning period, the parameters of the pre-trained encoder serve as the initialized weights for the segmentation encoder, and the whole framework is optimized by the segmentation loss $\mathcal{L}_{seg}$:

$$\mathcal{L}_{seg} = \mathbb{E}_{(x,y)\backsim(X,Y)}[-y \cdot logF(x)], \tag{6}$$

where $X$ and $Y$ are the input and ground truth set respectively, and $F$ is the fine-tuned segmentation model. During inference, not all points of the input mesh are passed to the network due to GPU memory constraints. Instead, we follow the approach used in previous research [17] and randomly sample 16,000 points. A KNN-based voting method is then employed to generate predictions for all vertices of the mesh.

## 3 Experiments

### 3.1 Experiment settings

**Datasets and evaluation** For self-supervised pre-training, we collected a large-scale IOS scan dataset comprising 6,000 unlabeled 3D scans of diverse tooth morphologies from patients in Hong Kong, China. For supervised fine-tuning, we employed the publicly available 3DTeethSeg dataset [2], containing 1,800 labeled lower and upper 3D IOS scans. Following [2], we split the fine-tuning dataset into three subsets: 1,000 samples for training, 200 for validation, and 600 for testing. All teeth were identified using the ISO-3950 notation system. To comprehensively evaluate both existing methods and our approach, we adopted three widely-used segmentation metrics: the Jaccard Index (mIoU), the Dice Similarity Coefficient (DSC), and the point-wise classification accuracy (Acc). All metrics were computed using standard definitions, with higher values indicating better performance.

**Implementation details** In self-supervised pre-training, we randomly sample $N$=16,000 points from each input 3D IOS scan and organize them into $M$=1,024 patches, each containing $K$=64 points, with 90% of patches randomly masked. The balancing factors $\lambda_1$, $\lambda_2$, and $\lambda_3$ are empirically set to 1.0, 0.1, and 0.001 based on validation results. For supervised fine-tuning, we maintain the same patch configuration ($M$=1,024) but without masking. Both stages utilize

**Table 1.** Results obtained by different approaches. † denotes methods pre-trained with the large-scale dataset. The top two results are marked as **bold** and underlined.

| Methods | Maxillary | | | Mandible | | | All | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc(%) | mIoU(%) | DSC(%) | Acc(%) | mIoU(%) | DSC(%) | Acc(%) | mIoU(%) | DSC(%) |
| **Supervised** | | | | | | | | | |
| PointNet++[20] | 86.56 | 77.55 | 85.21 | 82.55 | 75.23 | 82.52 | 84.54 | 76.02 | 83.52 |
| DGCNN[22] | 88.68 | 78.38 | 86.23 | 84.64 | 74.34 | 84.28 | 86.79 | 76.24 | 85.78 |
| MeshSegNet[13] | 88.25 | 79.64 | 87.14 | 85.62 | 76.93 | 85.79 | 86.29 | 78.30 | 86.12 |
| DC-Net [9] | 92.74 | 84.60 | 87.37 | 87.82 | 77.11 | 86.81 | 89.58 | 80.25 | 86.53 |
| GRAB-Net[15] | **92.86** | **86.13** | **90.53** | 89.15 | **82.62** | **88.70** | **91.61** | **83.79** | **89.30** |
| Transformer[25] | 92.53 | 85.87 | 90.28 | **89.69** | 81.89 | 87.85 | 91.11 | 83.38 | 89.07 |
| **Self-supervised** | | | | | | | | | |
| STSNet†[17] | 92.45 | 85.56 | 90.26 | 90.56 | 82.56 | 89.02 | 91.42 | 83.73 | 89.64 |
| PointBERT†[25] | 91.84 | 85.47 | 88.14 | 89.71 | 81.52 | 88.29 | 90.87 | 82.36 | 88.63 |
| PointMAE†[18] | 93.88 | 86.49 | 90.12 | 90.36 | 83.38 | 89.31 | 91.71 | 83.95 | 90.15 |
| Ours† | **95.19** | **88.36** | **93.24** | **92.95** | **86.35** | **91.61** | **94.07** | **87.35** | **92.42** |
| Ours w/o GAM | 94.37 | 87.93 | 91.39 | 90.92 | 84.93 | 90.22 | 92.78 | 85.12 | 90.82 |
| Ours w/o MGR | 94.67 | 87.55 | 92.43 | 91.38 | 85.56 | 90.53 | 92.44 | 86.35 | 91.57 |

AdamW optimizer with an initial learning rate of $5e^{-4}$ that decays to $5e^{-2}$ following cosine annealing. We employ a batch size of 2 and train for 100 and 200 epochs during pre-training and fine-tuning, respectively.

### 3.2 Main results

To demonstrate the effectiveness of our proposed method, we conducted comprehensive comparisons with several state-of-the-art segmentation approaches, including PointNet++ [20], DGCNN [22], MeshSegNet [13], DC-Net [9], and GRAB-Net [15]. The experimental results are presented in Table 1, where methods pre-trained on the large-scale dataset are denoted with †. Notably, our STEAM† achieves substantial improvements over its supervised counterpart Transformer [25], with gains of 2.96% in Acc and 3.97% in mIoU, validating the effectiveness of our pre-training strategy. Furthermore, our method outperforms the current state-of-the-art supervised method, GRAB-Net [15], by significant margins of 2.46% and 3.56% in Acc and mIoU, respectively. These remarkable results demonstrate that a vanilla transformer architecture, when properly pre-trained on large-scale data, can achieve superior performance without requiring sophisticated architectural designs or complex modifications.

Moreover, we compare our method with other self-supervised pre-training approaches, including the contrastive pre-training framework STSNet [17] specifically designed for tooth point cloud, and several generative pre-training methods designed for general point clouds, including PointBERT [25] and PointMAE [18]. For a fair comparison, all methods (except STSNet which uses customized data augmentations) employ identical augmentation techniques and the same Transformer [25] backbone. Following the setting described in Section 2.4, we transfer the pre-trained transformer encoder to the downstream segmentation task. As shown in Table 1, our STEAM† outperforms the best-performing pre-training

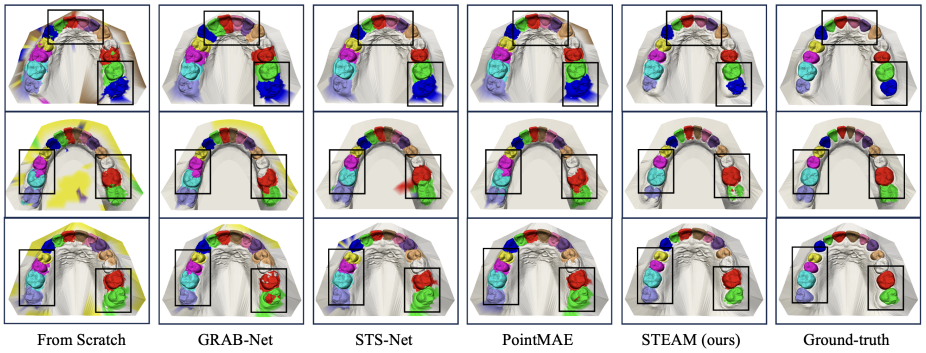


**Fig. 2.** Illustration of the segmentation results of previous works and ours.

method PointMAE† by 2.36% in Acc and 3.40% in mIoU, demonstrating superior knowledge learning from unlabeled tooth point clouds. Visual results in Fig. 1 show that compared to GRAB-Net [15] and other pre-training methods [17,18], STEAM† achieves better segmentation with clearer boundaries and can correctly identify teeth with similar shapes but different categories (black boxes in Fig. 2).

### 3.3 Ablation study

To thoroughly evaluate our proposed designs, we conducted extensive ablation studies as presented in Table 1. The results demonstrate that STEAM substantially outperforms the baseline PointMAE, achieving a 3.40% improvement in mIoU, which indicates its superior representation learning capability for downstream fine-tuning. To analyze the individual contributions of GAM and MGR, we conducted ablation experiments by removing each component. Results show that removing GAM leads to performance degradation of 1.29%, 2.23%, and 1.60% in Acc, mIoU, and DSC respectively, validating the effectiveness of hard patch selection for reconstruction. Similarly, the absence of MGR results in performance drops of 1.63%, 1.00%, and 0.85% in Acc, mIoU, and DSC, suggesting that reconstructing additional properties like surface geometries through MGR contributes positively to downstream performance.

## 4 Conclusion

In this paper, we present STEAM, the first masked pre-training framework for tooth point cloud segmentation, which leverages large-scale unlabeled data to reduce the dependency on labor-intensive annotations while improving model generalizability. To address the limitations of random masking and simple reconstruction targets in existing pre-training methods, we introduce GAM for selective hard patch reconstruction and MGR for multi-attribute geometry reconstruction. Extensive experimental results validate the effectiveness of our

framework and its individual components, offering a promising solution for practical tooth segmentation with large-scale datasets.

**Acknowledgments.** This work was supported by CUHK SSFCRS 23/24.

**Disclosure of Interests.** The authors have no competing interests to declare relevant to this article's content.

## References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Ben-Hamadou, A., Smaoui, O., Rekik, A., Pujades, S., Boyer, E., Lim, H., Kim, M., Lee, M., Chung, M., Shin, Y.G., et al.: 3dteethseg'22: 3d teeth scan segmentation and labeling challenge. arXiv preprint arXiv:2305.18277 (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NIPS **33**, 1877–1901 (2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020)
5. Cohen-Steiner, D., Morvan, J.M.: Restricted delaunay triangulations and normal cycle. In: Proceedings of the nineteenth annual symposium on Computational geometry. pp. 312–321 (2003)
6. Cui, Z., Li, C., Chen, N., Wei, G., Chen, R., Zhou, Y., Shen, D., Wang, W.: Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model. Med. Image Anal. **69**, 101949 (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. 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)
9. Hao, J., Liao, W., Zhang, Y., Peng, J., Zhao, Z., Chen, Z., Zhou, B., Feng, Y., Fang, B., Liu, Z., et al.: Toward clinically applicable 3-dimensional tooth segmentation via deep learning. J. Dent. Res. **101**(3), 304–311 (2022)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
12. Im, J., Kim, J.Y., Yu, H.S., Lee, K.J., Choi, S.H., Kim, J.H., Ahn, H.K., Cha, J.Y.: Accuracy and efficiency of automatic tooth segmentation in digital dental models using deep learning. Sci. Rep. **12**(1), 9429 (2022)
13. Lian, C., Wang, L., Wu, T.H., Wang, F., Yap, P.T., Ko, C.C., Shen, D.: Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners. IEEE Trans. Med. Imaging **39**(7), 2440–2450 (2020)
14. Liu, H., Cai, M., Lee, Y.J.: Masked discrimination for self-supervised learning on point clouds. In: ECCV. pp. 657–675. Springer (2022)

15. Liu, Y., Li, W., Liu, J., Chen, H., Yuan, Y.: Grab-net: Graph-based boundary-aware network for medical point cloud segmentation. IEEE Trans. Med. Imaging (2023)
16. Liu, Y., Li, W., Wang, C., Chen, H., Yuan, Y.: When 3d partial points meets sam: Tooth point cloud segmentation with sparse labels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 778–788. Springer (2024)
17. Liu, Z., He, X., Wang, H., Xiong, H., Zhang, Y., Wang, G., Hao, J., Feng, Y., Zhu, F., Hu, H.: Hierarchical self-supervised learning for 3d tooth segmentation in intra-oral mesh scans. IEEE Trans. Med. Imaging (2022)
18. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: ECCV. pp. 604–621. Springer (2022)
19. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
20. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NIPS **30** (2017)
21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
22. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. **38**(5), 1–12 (2019)
23. Xiong, H., Li, K., Tan, K., Feng, Y., Zhou, J.T., Hao, J., Ying, H., Wu, J., Liu, Z.: Tsegformer: 3d tooth segmentation in intraoral scans with geometry guided transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 421–432. Springer (2023)
24. Xu, X., Liu, C., Zheng, Y.: 3d tooth segmentation and labeling using deep convolutional neural networks. IEEE Trans. Vis. Comput. Graph. **25**(7), 2336–2348 (2018)
25. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: CVPR. pp. 19313–19322 (2022)
26. Yuan, T., Liao, W., Dai, N., Cheng, X., Yu, Q.: Single-tooth modeling for 3d dental model. Int. J. Biomed. Imaging **2010** (2010)
27. Zanjani, F.G., Moin, D.A., Claessen, F., Cherici, T., Parinussa, S., Pourtaherian, A., Zinger, S., de With, P.H.: Mask-mcnet: Instance segmentation in 3d point cloud of intra-oral scans. In: MICCAI. pp. 128–136. Springer (2019)
28. Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: ICCV. pp. 10252–10263 (2021)