

3DTeethSAM: Taming SAM2 for 3D Teeth Segmentation

Zhiguo Lu¹, Jianwen Lou^{1*}, Mingjun Ma², Hairong Jin³, Youyi Zheng³, Kun Zhou³

¹School of Software Technology, Zhejiang University

²College of Intelligence and Computing, Tianjin University

³State Key Lab of CAD&CG, Zhejiang University
{lzg, jianwen.lou}@zju.edu.cn

Abstract

3D teeth segmentation, involving the localization of tooth instances and their semantic categorization in 3D dental models, is a critical yet challenging task in digital dentistry due to the complexity of real-world dentition. In this paper, we propose 3DTeethSAM, an adaptation of the Segment Anything Model 2 (SAM2) for 3D teeth segmentation. SAM2 is a pretrained foundation model for image and video segmentation, demonstrating a strong backbone in various downstream scenarios. To adapt SAM2 for 3D teeth data, we render images of 3D teeth models from predefined views, apply SAM2 for 2D segmentation, and reconstruct 3D results using 2D-3D projections. Since SAM2’s performance depends on input prompts and its initial outputs often have deficiencies, and given its class-agnostic nature, we introduce three lightweight learnable modules: (1) a prompt embedding generator to derive prompt embeddings from image embeddings for accurate mask decoding, (2) a mask refiner to enhance SAM2’s initial segmentation results, and (3) a mask classifier to categorize the generated masks. Additionally, we incorporate Deformable Global Attention Plugins (DGAP) into SAM2’s image encoder. The DGAP enhances both the segmentation accuracy and the speed of the training process. Our method has been validated on the 3DTeethSeg benchmark, achieving an IoU of 91.90% on high-resolution 3D teeth meshes, establishing a new state-of-the-art in the field.

Code — <https://github.com/Crisitofy/3DTeethSAM>

Introduction

Analyzing 3D dental models, which provide accurate and high-resolution representations of a patient’s oral anatomy, is a fundamental aspect of digital dentistry. This analysis enables precise diagnosis, effective treatment planning, and the creation of customized dental solutions. A crucial step in this process is the segmentation and classification of individual teeth, which facilitates further tasks such as orthodontic staging. However, 3D teeth segmentation remains a challenging problem due to two primary issues: (1) the complexity of real-world dentition, including anatomical variations and dental anomalies such as crowding and crooked teeth, and (2) the scarcity of labeled data.

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Current methods for 3D teeth segmentation predominantly rely on deep neural networks (e.g., PointNet++(Qi et al. 2017b), TSGCNet(Zhao et al. 2021b), and TSRNet (Jin et al. 2024)) that operate directly on 3D dental models represented as meshes or point clouds. While these methods show promising results on low-resolution models, they struggle to scale to high-resolution 3D ones due to the limited capacity of the networks. In contrast, 2D vision foundation models, such as the Segment Anything Model 2 (SAM2) (Ravi et al. 2024), trained on billions of 2D masks, have demonstrated impressive image segmentation performance and generalization capabilities across various tasks. This success suggests a potential pathway for leveraging SAM2’s pretrained capabilities in 3D teeth segmentation. However, adapting SAM2 for this task poses significant challenges, including the inherent 2D-3D dimensionality mismatch, the need for manual prompts (such as points or boxes) to guide segmentation, and the difficulty of fine-tuning SAM2 effectively while preserving its pretrained weights.

In this paper, we propose 3DTeethSAM, a novel framework that adapts SAM2 for automatic 3D teeth segmentation. As shown in Figure 1, our approach starts by rendering the 3D dental mesh into multiple 2D images from predefined viewing angles. These images are then processed by a customized version of SAM2 to generate segmentation masks. The 2D segmentation results are lifted back into 3D space using a voting strategy, which aggregates information from multiple views. To further refine the segmentation, we apply Graph Cut (Boykov and Jolly 2001) to correct boundary inaccuracies. A key innovation of our approach is the adaptation of SAM2 to 3D teeth segmentation through lightweight adapters and an image embedding enhancement scheme. These adapters, including a Prompt Embedding Generator for generating informative prompt embeddings, a Mask Refiner to improve the coarse segmentation, and a Mask Classifier to enable semantic label recognition, enhance SAM2’s ability to handle 2D teeth images effectively. Additionally, we introduce the Deformable Global Attention Plugin (DGAP) into SAM2’s image encoder. DGAP dynamically samples features during global attention, helping the model focus on the region of interest, i.e., the teeth. Extensive validation on the Teeth3DS benchmark demonstrates that 3DTeethSAM achieves a mean IoU of 91.90%, outperforming existing methods and establishing an effective

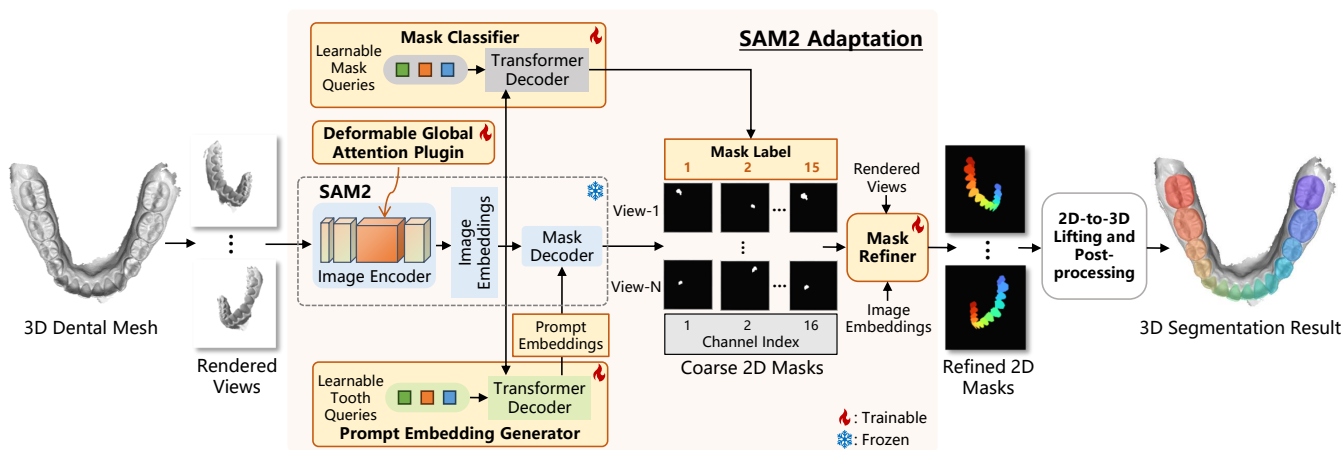


Figure 1: Pipeline of 3DTeethSAM. 3D dental models are rendered into 2D images from predefined views and processed by an adapted SAM2 for segmentation. The 2D results are then reconstructed in 3D. To enhance performance, three lightweight modules are introduced: (1) Prompt Embedding Generator for mask decoding, (2) Mask Refiner for segmentation improvement, and (3) Mask Classifier for semantic labeling. Additionally, the Deformable Global Attention Plugin (DGAP) refines feature sampling in SAM2’s image encoder.

framework for adapting 2D foundation models to complex 3D segmentation tasks. The main contributions can be summarized as follows:

- We propose 3DTeethSAM, a novel method that adapts the Segment Anything Model 2 (SAM2) for 3D teeth segmentation. The method includes three lightweight, learnable modules that enable auto-prompting and improve SAM2’s initial outputs, ensuring accurate segmentation and categorization of teeth. Additionally, we incorporate Deformable Global Attention Plugins (DGAP) to enhance both segmentation accuracy and training efficiency.
- We conduct extensive experiments that demonstrate the effectiveness of the proposed method and provide valuable insights into its key components.

Related Work

3D Dental Segmentation

The field of 3D dental segmentation has evolved from traditional geometry-based methods (Yuan et al. 2010; Wu et al. 2014; Zou et al. 2015) to deep learning approaches. Early techniques, which relied on curvature or watershed algorithms (Li, Ning, and Wang 2007), set initial benchmarks but struggled with the geometric complexity of real-world dentitions. The advent of deep learning, driven by networks like PointNet (Qi et al. 2017a) and DGCNN (Wang et al. 2019), led to the development of powerful, data-driven solutions. Today, state-of-the-art methods focus on specialized architectures, such as those that learn from mesh structures (e.g., MeshSegNet (Lian et al. 2020)), incorporate transformers for geometric context (e.g., TsegFormer (Xiong et al. 2023)), or perform post-hoc refinement (Jin et al. 2024). A common limitation among these advanced methods is their reliance on standalone networks trained from

scratch on domain-specific data, which prevents them from leveraging the vast knowledge encapsulated in large-scale pre-trained models.

Segment Anything Model

The advent of 2D vision foundation models, particularly the Segment Anything Model (SAM) (Kirillov et al. 2023) and its successor SAM2 (Ravi et al. 2024), has revolutionized computer vision with impressive zero-shot, prompt-based image segmentation. This has led to various adaptations, such as MedSAM (Ma et al. 2024), which tailors SAM for medical imagery, and efforts (Ke et al. 2023; Fan et al. 2023) to enhance its adaptability to downstream tasks. While SAM2 has proven highly successful in 2D image segmentation, its application to 3D segmentation remains largely unexplored. Key challenges include the inherent 2D-3D dimensionality mismatch, reliance on manual prompts (e.g., points or boxes) for guidance, and difficulties in fine-tuning SAM2 while preserving its pretrained weights. Despite these challenges, applying 2D foundation models like SAM2 to 3D teeth segmentation holds great promise, a direction we explore in this study.

Methodology

We propose 3DTeethSAM, a novel framework that leverages the 2D vision foundation model SAM2 (Ravi et al. 2024) for 3D teeth segmentation. As illustrated in Figure 1, our approach begins by rendering a 3D dental mesh into multiple images from predefined viewing angles. These images are then processed using a customized version of SAM2 to perform segmentation. The resulting 2D segmentation masks are lifted back into 3D space via an intuitive voting strategy, which aggregates information across different views. Finally, to refine the segmentation, we apply Graph Cut (Boykov and Jolly 2001) to correct issues like inac-

curacies in the boundaries. At the heart of our approach is the adaptation of a pretrained SAM2 model to effectively handle teeth images. This is achieved by introducing learnable lightweight adapters and an image embedding enhancement scheme, while preserving SAM2’s pretrained weights. Specifically, the adapters include a Prompt Embedding Generator, which creates informative prompt embeddings from image features for mask decoding; a Mask Refiner, which optimizes the coarse segmentation results produced by SAM2; and a Mask Classifier, which enables SAM2 to be class-aware by recognizing semantic labels in the segmentation map. To extract an enhanced image embedding for the adapters, we integrate a key component into SAM2’s image encoder: Deformable Global Attention Plugin (DGAP). DGAP dynamically samples features during global attention, helping the model focus on the region of interest, i.e., the teeth. In the following sections, we first introduce the preliminaries, including the basic concepts of 3D teeth segmentation and SAM2, and then describe our proposed method in detail.

Preliminaries

3D Teeth Segmentation Both the upper and lower halves of an adult’s jaw contain up to 16 teeth, with each tooth having a unique ID. The goal of 3D teeth segmentation is to assign a class label to each vertex in the 3D dental model, identifying one of the 16 teeth or the background. In this study, a 3D dental model is represented as a 3D mesh of either the upper or lower teeth. It is denoted as $\mathcal{M}(P, F)$, where $P = \{p_n \in \mathbb{R}^3, n = 1, 2, \dots, N\}$ is the set of 3D points, and $F = \{(p_i, p_j, p_k)_m, m = 1, 2, \dots, M\}$ is the set of triangles that define the connectivity between points.

SAM2 Segment Anything Model 2 (SAM2) is a groundbreaking vision foundation model designed for promptable segmentation in images and videos. It consists of three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder, based on the Vision Transformer (ViT) architecture, processes the input image and generates an embedding that captures high-level visual features. The prompt encoder embeds prompts, such as points, boxes, or masks. The mask decoder combines the above two information sources to predict segmentation masks.

Multi-View Teeth Image Rendering

Given a 3D dental mesh \mathcal{M} , we render it into multi-view images to align with SAM2 for segmentation. This process is designed to comprehensively capture both structural and surface information of the 3D dental mesh from a set of carefully chosen viewpoints, ensuring that critical features for segmentation are visible in at least one rendered view. Specifically, each 3D dental mesh is first normalized by translating its center to the origin of a uniform coordinate system and rotating it so that the dental crown is oriented upwards. The transformed mesh is then rendered into 512x512 RGB images from a set of fixed camera viewpoints (see Figure 1), including a frontal view, a back view, and several side

views. The rendering process is formulated as follows:

$$I_v = \Pi_v(\mathcal{M}) \quad (1)$$

where $I_v \in \mathbb{R}^{3 \times 512 \times 512}$ represents the rendered teeth image, and Π_v denotes the camera projection from the viewing angle v . The number of views is empirically set to balance segmentation accuracy and computational complexity.

Using the above equation, we can also render a teeth mask map $Mask_v \in \mathbb{R}^{16 \times 512 \times 512}$ from the 3D dental mesh based on point-wise segmentation labels. The mask map has 16 channels, each corresponding to one of the 16 teeth. Its pixel values are normalized to the range 0 to 1 along each channel. The value indicates the probability that a pixel belongs to the tooth category associated with that particular channel.

Teeth Image Segmentation via SAM2 Adaptation

Trained on a large-scale dataset of natural images, SAM2 demonstrates strong zero-shot learning performance across a wide range of downstream tasks. However, applying the pretrained SAM2 model to teeth images is not straightforward. There are three key challenges: first, SAM2 requires input prompts for segmentation; second, the original segmentation results produced by SAM2 show noticeable deficiencies; and third, SAM2 is class-agnostic. These issues prevent the vanilla SAM2 model from meeting the automation and precision requirements for teeth image segmentation. To leverage the pre-learned knowledge of SAM2 and adapt it for teeth image segmentation in a fully automatic and parameter-efficient manner, we introduce three lightweight SAM2 adapters: a Prompt Embedding Generator, a Mask Refiner, and a Mask Classifier. To further enhance the model’s adaptation, we propose integrating the Deformable Global Attention Plugin (DGAP) into SAM2’s image encoder, which improves feature extraction for the adapters. During training, we keep SAM2’s pretrained weights frozen, while only optimizing those of the adapters and DGAP.

Prompt Embedding Generator SAM2 uses prompts such as points, boxes, and masks to identify the target image region for segmentation. The type and location of the prompts are critical for accurate segmentation, but they are often difficult to determine. Additionally, when multiple segmentation targets are present in an image and have an inherent structure (e.g., teeth arranged in a canonical pattern), it becomes essential to model the positional relationships between the prompts. To address these challenges, we propose a trainable generator capable of predicting prompt embeddings directly from image features. Specifically, inspired by DETR (Carion et al. 2020), we employ a Transformer decoder for this generation process. The decoder follows the architecture of the original Transformer decoder introduced in (Vaswani et al. 2017). It accepts randomly initialized query vectors as input and transforms them into prompt embeddings for tooth instances using multi-layer self-attention and cross-attention. The self-attention module models pairwise relationships between the queries, while the cross-attention module aligns the queries with the image features. This approach allows the model to reason about all prompt embeddings simultaneously, leveraging both the

relationships between tooth instances and the broader image context. Given that a teeth image can contain up to 16 teeth, we set the number of query vectors to 16. To handle cases with missing teeth, we also learn a confidence score from each prompt embedding, using a fully connected layer and a Sigmoid function. This score indicates the validity of each prompt embedding, with values ranging from 0 to 1. A higher score represents a greater probability that the corresponding tooth instance exists.

Mask Refiner Using the generated prompt embeddings and the teeth image embedding output by SAM2’s image encoder, we can create a 16-channel mask map with SAM2’s mask decoder. Each channel in this map corresponds to the segmentation mask of a specific tooth. However, these masks may not precisely localize tooth instances, particularly along the boundaries, due to SAM2’s general-domain pre-training. To address the limitations of the coarse masks, we introduce a mask refiner, a specialized convolutional neural network designed to enhance boundary precision. The refiner processes three key inputs:

- Teeth image: Provides rich, low-level shape and texture details crucial for precise boundary delineation.
- Coarse mask map: Offers strong spatial priors on the tooth instances’ location and shape.
- SAM2’s image embedding: Captures high-level semantic context.

The backbone of the refiner is based on a UNet (Ronneberger, Fischer, and Brox 2015) architecture, which is commonly used for image segmentation. In the contracting path of the UNet, each convolutional layer contains three main streams, with each stream transforming a specific input into latent representations. These latent features are then concatenated and passed to the subsequent convolutional layer and the expansive path of the UNet via a skip connection.

Mask Classifier The original SAM2 model does not support identifying the semantic category of each segmentation mask. To address this limitation, a straightforward approach would be to bind the channels of the mask map $Mask_v \in \mathbb{R}^{16 \times 512 \times 512}$ to 16 tooth IDs in a one-to-one manner. In this setup, each channel would correspond to a specific tooth ID, allowing the model to infer the tooth identity directly from the channel index. However, our experiments show that this naive method is prone to channel-to-ID mismatches, often assigning a tooth mask to the wrong channel, especially in cases with missing teeth. To overcome this issue, we introduce a mask classifier that explicitly identifies the tooth ID associated with each channel in the mask map. The classification process resembles the prompt embedding generation in SAM2: it requires capturing both the spatial correlation between teeth (implied by their natural arrangement) and the image context. To this end, we adopt the same architecture as the prompt embedding generator, using a Transformer decoder (Vaswani et al. 2017) to transform 16 randomly initialized query vectors (each representing a potential tooth instance) into 16 class probability vectors, based on image features. The only architectural difference is

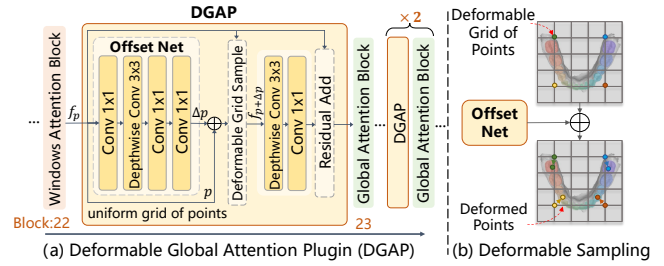


Figure 2: Illustration of the proposed Deformable Global Attention Plugin (DGAP). (a) Architecture of the DGAP module. (b) Deformable sampling grid based on the Offset Net.

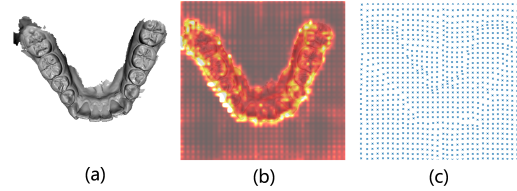


Figure 3: Effect of the Deformable Global Attention Plugin (DGAP). (a) Rendered teeth image. (b) Attention heatmap after applying DGAP. (c) Deformed sampling points that align with the shape of the teeth.

in the classifier’s final layers, which consist of a multi-layer perceptron (MLP) followed by a softmax activation. To handle missing teeth, we extend each class probability vector with an additional dimension representing the background. This allows the model to classify absent tooth instances as background, ensuring robustness in cases with incomplete dental data.

Deformable Global Attention Plugin The above three adapters all require an image embedding as input for prediction. Although the feature pyramid produced by SAM2’s image encoder offers a handy image embedding for the adapters, we argue that such a feature representation is designed for SAM2’s mask decoding, not purely fitting with the adaptation process. A simple solution to this problem would be full-parameter tuning SAM2’s image encoder during adaptation, which however is very costly. As a remedy, we propose to integrate a lightweight Deformable Global Attention Plugin into SAM2’s image encoder for capturing informative features for adapters in a parameter-efficient way (see Figure 2). The plugin is inspired by the deformable attention mechanism introduced in (Xia et al. 2022). Deformable attention uses a dynamic feature map to generate key and value embeddings. It learns an offset network that predicts a set of offsets from the query embedding, which are then used to deform a grid of reference points on the feature map. This process enables dynamic feature extraction during the attention operation. The deformed points are shown to be concentrated around the target region in the image, which leads to more informative feature extraction. We incorporate deformable attention into SAM2’s image encoder with a systematic adaptation as follows:

- **Integration into Global Attention Block:** Deformable attention is integrated into each global attention block in the 3rd stage of SAM2’s image encoder, specifically in the Hiera trunk. The Hiera trunk consists of four main stages, each containing multiple attention blocks. As the feature map down-samples across the stages, it becomes progressively more abstract. The 3rd stage is crucial as it contains the majority of attention blocks, which are essential for learning image embeddings. Therefore, we focus on deformable attention in this stage to extract the most meaningful features. Additionally, we choose the global attention block to maintain a broad receptive field.
- **Embedding Prediction from Deformed Feature Map:** We predict not only the key and value embeddings but also the query embedding from the deformed feature map. This approach offers more flexibility than the standard deformable attention method. It also results in a plug-and-play module, where the offset network and dynamic sampling can be added before the global attention operation without altering the internal implementation.
- **Skip Connection for Feature Map Combination:** We combine the deformed and undeformed feature maps using a skip connection, allowing the model to leverage both, which results in more robust feature learning.

Training Losses Our training employs a composite loss function to address SAM2’s adaptation challenges: automated prompting, semantic classification, and boundary refinement. We use the Hungarian algorithm (Kuhn 1955) for optimal one-to-one assignment between predicted queries and ground-truth annotations, enabling instance-aware supervision.

The total loss function jointly optimizes our three lightweight adapters:

$$L_{\text{total}} = \lambda_{\text{MC}}L_{\text{MC}} + \lambda_{\text{PEG}}L_{\text{PEG}} + \lambda_{\text{MR}}L_{\text{MR}} \quad (2)$$

where λ_{MC} , λ_{PEG} , and λ_{MR} are hyperparameter weights that balance the contribution of each adapter.

Mask Classifier Loss (L_{MC}): The classifier addresses SAM2’s class-agnostic limitation by learning to distinguish between different tooth types. We apply Cross-Entropy loss over 17 classes (16 teeth + background) to matched query-target pairs:

$$L_{\text{MC}} = - \sum_{i \in \mathcal{M}} \log p_i(c_i^*) \quad (3)$$

where \mathcal{M} denotes matched pairs, c_i^* is the ground-truth class, and $p_i(c_i^*)$ is the predicted probability.

PEG Loss (L_{PEG}): The PEG must generate high-quality prompt embeddings that produce accurate initial masks while handling missing teeth scenarios. Its loss combines three complementary terms:

$$L_{\text{PEG}} = \lambda_{\text{bce}}L_{\text{bce}} + \lambda_{\text{dice}}L_{\text{dice}} + \lambda_{\text{conf}}L_{\text{conf}} \quad (4)$$

Here, L_{bce} provides pixel-wise supervision through binary cross-entropy, L_{dice} optimizes region-level overlap to handle class imbalance, and L_{conf} trains confidence scores to predict tooth presence, enabling robust handling of incomplete dental arches.

Mask Refiner Loss (L_{MR}): The refiner transforms coarse masks into precise segmentations with sharp boundaries—critical for clinical applications. Its multi-objective loss ensures both semantic accuracy and boundary precision:

$$L_{\text{MR}} = \lambda_{\text{ce}}L_{\text{ce}} + \lambda_{\text{dice}}L_{\text{dice}} + \lambda_{\text{boundary}}L_{\text{boundary}} \quad (5)$$

where L_{ce} applies multi-class Cross-Entropy over all 17 classes, L_{dice} maximizes regional overlap, and L_{boundary} measures L1 distance between spatial gradients of predicted and ground-truth masks using Sobel filters, encouraging sharp tooth-gingiva boundaries essential for treatment planning.

2D-to-3D Segmentation Lifting and Postprocessing

After obtaining the 2D segmentation results using the adapted SAM2 model, we lift these results into 3D space. Specifically, we invert the projection matrix applied during multi-view image rendering and assign the segmentation label of each image pixel to its corresponding 3D mesh vertex. In cases where multiple 2D segmentation labels are associated with a single mesh vertex, we select the label that appears most frequently across all rendered views. As in previous studies, we apply the well-known Graph Cut approach (Boykov and Jolly 2001) to further refine the 3D segmentation by filling holes and improving boundary precision.

Experiments

Datasets We conduct comprehensive experiments on the Teeth3DS benchmark, the most challenging publicly available dataset for 3D dental segmentation. The dataset comprises 1,800 high-resolution intraoral 3D scans from 900 patients, with both upper and lower dental arches included. Following the FDI dental notation system, each dental arch contains up to 16 tooth instances plus gingival background, resulting in 17 semantic classes. We strictly adhere to the official train/test split (1,200/600 scans) to ensure fair comparison with prior methods.

Implementation Details All experiments are conducted using PyTorch on NVIDIA RTX 4090 GPUs. We employ SAM2 (Hiera-L) as our foundation model backbone, keeping its pre-trained weights frozen. The network processes original-resolution 3D meshes without downsampling, with multi-view renderings dynamically generated at 512×512 resolution.

Training adopts an end-to-end scheme with AdamW optimizer (learning rate: 2e-4, cosine annealing schedule with 5-epoch warm-up). Models are trained for 100 epochs with batch size 4, utilizing mixed-precision training for efficiency. The loss weights are empirically set as: $\lambda_{\text{MC}} = 1.0$, $\lambda_{\text{PEG}} = 1.0$, $\lambda_{\text{MR}} = 2.0$, with sub-loss weights following standard practices.

Evaluation Metrics The evaluation encompasses four complementary metrics: overall accuracy (OA), tooth-wise mIoU (T-mIoU), boundary IoU (B-IoU), and Dice score. OA measures per-vertex classification correctness across the entire dental mesh. T-mIoU computes the mean IoU across all individual tooth instances, providing an instance-level assessment of segmentation quality. The Dice score

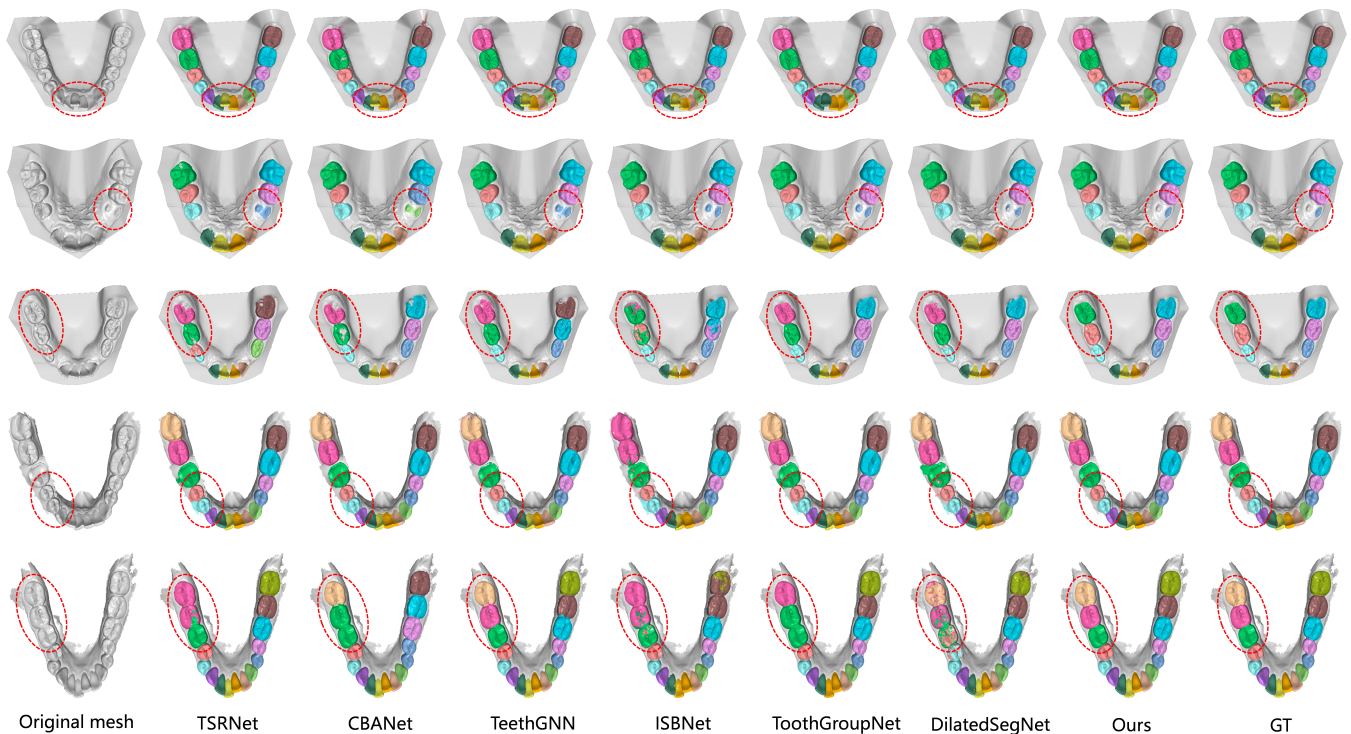


Figure 4: Visual comparison of different methods.

offers a global measure of region overlap between predicted and ground-truth masks. B-IoU specifically evaluates boundary precision by focusing on vertices near inter-tooth boundaries, where a vertex is considered a boundary point if there exist vertices with different labels within its k -neighbourhood ($k = 10$). This boundary-focused metric is particularly crucial for accurate crown–gingiva delineation in clinical applications.

Comparison on Teeth3DS Table 1 presents quantitative comparisons against 11 state-of-the-art methods spanning different paradigms: point cloud networks (PointNet++, DGCNN), mesh-based approaches (MeshSegNet, iMeshSegNet), graph neural networks (TeethGNN, TSGCNet), and recent advanced methods (TSRNet, ToothGroupNet).

3DTeethSAM achieves state-of-the-art performance across all metrics: 95.48% OA, 91.90% T-mIoU, 70.05% B-IoU, and 94.33% Dice. Notably, our method demonstrates substantial improvements over the best-performing prior method (ToothGroupNet) by +1.74% T-mIoU and +0.75% B-IoU, establishing new benchmarks while using fewer trainable parameters than methods trained from scratch.

Looking at the eight symmetric tooth groups in the right-most columns, 3DTeethSAM consistently achieves the highest mIoU for every group, with particularly strong performance on rare and challenging categories. For instance, it attains 83.29% mIoU on wisdom teeth ($T_{8/16}$), significantly outperforming ToothGroupNet’s 68.2%. This 15.09-point improvement highlights the benefits of leveraging large-scale 2D pretraining, demonstrating how foundation models can effectively mitigate data scarcity issues that often hinder

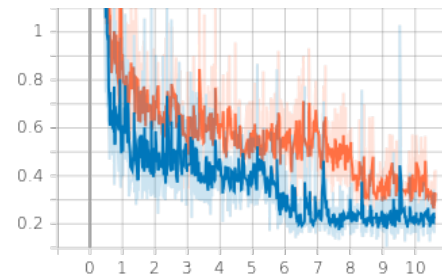


Figure 5: Training loss curves of our method on the Teeth3DS dataset. The blue curve (with DGAP) converges faster and reaches a lower final value compared to the orange curve (without DGAP), demonstrating the efficiency of the proposed DGAP.

specialized 3D architectures.

Figure 4 illustrates our method’s robustness across diverse clinical scenarios: dental crowding, tooth eruption, edentulous regions, gingival complexity, and complete arch segmentation. Unlike prior methods that suffer from boundary ambiguity or class confusion, 3DTeethSAM maintains precise inter-tooth boundaries and avoids common failure modes such as gingival leakage or adjacent tooth merging.

Ablation Studies To quantify the contribution of each module, we conduct an ablation study on the four key components of 3DTeethSAM. Starting from the full model, we progressively remove each component to evaluate its individual impact. The results are summarized in Table 2.

Method	OA	T _{all}	B _{all}	Dice	T _{1/9}	T _{2/10}	T _{3/11}	T _{4/12}	T _{5/13}	T _{6/14}	T _{7/15}	T _{8/16}
iMeshSegNet (Wu et al. 2022)	82.45	67.65	26.31	77.58	71.63	70.90	68.20	71.67	66.98	67.66	55.65	00.00
TSegNet (Cui et al. 2021)	78.22	59.81	28.00	67.35	57.05	62.89	69.03	58.74	52.23	61.45	66.02	00.00
TeethGNN (Zheng et al. 2022)	90.96	83.89	48.49	88.46	86.58	86.31	86.56	87.69	82.05	79.21	82.89	66.54
TSRNet (Jin et al. 2024)	92.87	86.56	51.11	90.91	88.20	87.83	87.35	87.65	86.57	86.69	83.74	70.82
ToothGroupNet (Ben-Hamadou et al. 2023)	95.19	90.16	69.30	92.88	92.19	92.30	92.65	93.44	87.74	88.84	84.82	68.2
TSGCNet (Zhang et al. 2021)	89.84	79.79	36.98	86.73	82.63	81.59	82.82	82.75	80.04	80.60	69.27	34.23
IsbNet (Ngo, Hua, and Nguyen 2023)	91.53	81.01	39.61	87.65	68.61	80.62	84.06	86.44	84.61	85.79	80.44	26.75
DGCNN (Wang et al. 2019)	90.21	80.08	28.60	87.27	80.85	80.31	81.12	82.67	79.86	81.53	75.45	52.42
CBAnet (Jin et al. 2025)	92.81	86.77	50.81	91.16	88.62	87.94	88.17	88.41	86.39	86.12	83.05	76.88
PT (Zhao et al. 2021a)	86.15	71.52	29.22	78.84	71.26	72.15	72.08	72.50	69.55	73.27	74.88	2.27
DilatedSegNet (Krenmayr et al. 2024)	93.40	86.55	51.57	91.26	86.49	86.46	87.33	88.61	86.83	88.09	83.88	62.49
Ours	95.48	91.90	70.05	94.33	93.28	93.36	93.16	93.79	91.65	90.73	89.10	83.29

Table 1: Comparison of segmentation performance. T_{all} and B_{all} are abbreviations for the mean IoU over all tooth classes (T-mIoU) and the boundary (B-IoU), respectively. Columns T_{1/9} through T_{8/16} present the per-tooth T-mIoU scores, with labels grouped for symmetric positions (e.g., T1 and T9).

Setting	OA \uparrow	T-mIoU \uparrow	B-IoU \uparrow	Dice \uparrow
Full Model	95.48	91.90	70.05	94.33
w/o DGAP	94.87	90.61	66.64	93.45
w/o PEG	76.52	52.46	28.17	58.88
w/o Mask Refiner	95.13	91.10	68.43	93.67
w/o Mask Classifier	95.21	91.31	67.56	93.98

Table 2: Ablation study on each proposed component. Removing any module results in a noticeable drop across all metrics, confirming the contribution of each module.

Prompt Embedding Generator (PEG): To evaluate PEG’s contribution, we replace it with manual prompts derived from ground-truth mask center points, simulating an idealized manual prompting scenario. Even with this oracle-level prompting, performance drops dramatically by 39.44% T-mIoU (91.90%→52.46%), demonstrating that PEG’s automated, instance-aware prompt generation is fundamentally superior to traditional point-based prompting. This massive degradation confirms that our learned prompt embeddings capture complex spatial relationships and contextual information that simple point prompts cannot provide.

Deformable Global Attention Plugin (DGAP): Removing DGAP degrades T-mIoU by 1.29% (91.90%→90.61%) and B-IoU by 3.41%, confirming the benefit of morphology-aware attention for dental structures. Beyond performance gains, Figure 5 illustrates that DGAP significantly accelerates training convergence, providing both accuracy and efficiency benefits through deformable feature sampling.

Mask Refiner: Removing the refiner reduces T-mIoU by 0.80% while degrading B-IoU by 1.62%, indicating its role in both overall segmentation quality and boundary precision.

Mask Classifier: Disabling the classifier decreases T-mIoU by 0.59% and B-IoU by 2.49%, demonstrating that explicit semantic reasoning provides measurable benefits. Figure 6 shows qualitative examples where the classifier correctly resolves tooth category assignments, particularly for morphologically similar adjacent teeth.

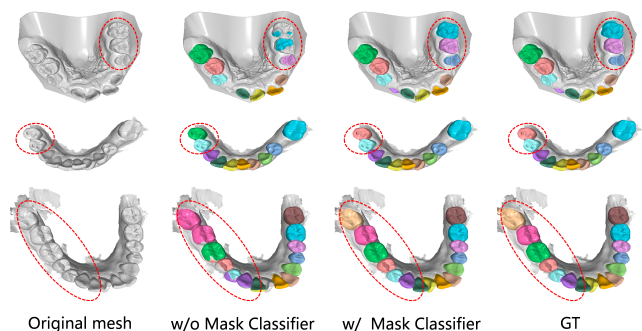


Figure 6: Visual results of the ablation study on the classifier module. The red ellipses highlight regions where the classifier corrects tooth category assignments, demonstrating its effectiveness in disambiguating adjacent teeth.

These findings confirm that each proposed component contributes complementary capabilities (DGAP for enhanced feature extraction, PEG for automated prompting, Mask Refiner for boundary refinement, and Mask Classifier for semantic disambiguation), which are essential for achieving SOTA 3D teeth segmentation performance.

Conclusion

In this paper, we present 3DTeethSAM, a novel framework that adapts SAM2 for automatic 3D teeth segmentation. Our approach involves rendering 3D dental meshes into 2D images, processing them with a customized SAM2 to generate segmentation masks, and reconstructing the results in 3D using a voting strategy. Key innovations include lightweight adapters for prompt generation, mask refinement, and semantic classification, as well as the introduction of the Deformable Global Attention Plugin (DGAP) to enhance feature sampling. Extensive validation on the Teeth3DS benchmark shows that 3DTeethSAM surpasses existing methods and demonstrating the effectiveness of adapting 2D foundation models for 3D segmentation tasks.

References

- Ben-Hamadou, A.; Smaoui, O.; Rekik, A.; Pujades, S.; Boyer, E.; Lim, H.; Kim, M.; Lee, M.; Chung, M.; Shin, Y.-G.; et al. 2023. 3DTeethSeg'22: 3D Teeth Scan Segmentation and Labeling Challenge. *arXiv preprint arXiv:2305.18277*.
- Boykov, Y. Y.; and Jolly, M.-P. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, 105–112. IEEE.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.
- Cui, Z.; Li, C.; Chen, N.; Wei, G.; Chen, R.; Zhou, Y.; Shen, D.; and Wang, W. 2021. TSegNet: An efficient and accurate tooth segmentation network on 3D dental model. *Medical Image Analysis*, 69: 101949.
- Fan, Q.; Tao, X.; Ke, L.; Ye, M.; Zhang, Y.; Wan, P.; Wang, Z.; Tai, Y.-W.; and Tang, C.-K. 2023. Stable segment anything model. *arXiv preprint arXiv:2311.15776*.
- Jin, H.; Lou, J.; Lu, Z.; Wu, T.; Zhou, K.; and Zheng, Y. 2025. Learning center-and boundary-aware instance representation for 3D tooth segmentation. *Computers & Graphics*, 104313.
- Jin, H.; Shen, Y.; Lou, J.; Zhou, K.; and Zheng, Y. 2024. TSRNet: A Dual-Stream Network for Refining 3D Tooth Segmentation. *IEEE Transactions on Visualization and Computer Graphics*.
- Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2023. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.
- Krenmayr, L.; von Schwerin, R.; Schaudt, D.; Riedel, P.; and Hafner, A. 2024. Dilatedtoothsegnet: Tooth segmentation network on 3d dental meshes through increasing receptive vision. *Journal of Imaging Informatics in Medicine*, 37(4): 1846–1862.
- Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.
- Li, Z.; Ning, X.; and Wang, Z. 2007. A fast segmentation method for stl teeth model. In *2007 IEEE/ICME International Conference on Complex Medical Engineering*, 163–166. IEEE.
- Lian, C.; Wang, L.; Wu, T.-H.; Wang, F.; Yap, P.-T.; Ko, C.-C.; and Shen, D. 2020. Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3D intraoral scanners. *IEEE transactions on medical imaging*, 39(7): 2440–2450.
- Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.
- Ngo, T. D.; Hua, B.-S.; and Nguyen, K. 2023. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13550–13559.
- Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.
- Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.
- Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.
- Wu, K.; Chen, L.; Li, J.; and Zhou, Y. 2014. Tooth segmentation on dental meshes using morphologic skeleton. *Computers & Graphics*, 38: 199–211.
- Wu, T.-H.; Lian, C.; Lee, S.; Pastewait, M.; Piers, C.; Liu, J.; Wang, F.; Wang, L.; Chiu, C.-Y.; Wang, W.; et al. 2022. Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3D intraoral scans. *IEEE transactions on medical imaging*, 41(11): 3158–3166.
- Xia, Z.; Pan, X.; Song, S.; Li, L. E.; and Huang, G. 2022. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4794–4803.
- Xiong, H.; Li, K.; Tan, K.; Feng, Y.; Zhou, J. T.; Hao, J.; Ying, H.; Wu, J.; and Liu, Z. 2023. Tsegformer: 3d tooth segmentation in intraoral scans with geometry guided transformer. In *International conference on medical image computing and computer-assisted intervention*, 421–432. Springer.
- Yuan, T.; Liao, W.; Dai, N.; Cheng, X.; and Yu, Q. 2010. Single-tooth modeling for 3D dental model. *International journal of biomedical imaging*, 2010(1): 535329.
- Zhang, L.; Zhao, Y.; Meng, D.; Cui, Z.; Gao, C.; Gao, X.; Lian, C.; and Shen, D. 2021. TSGCNet: Discriminative geometric feature learning with two-stream graph convolutional

network for 3D dental model segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6699–6708.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021a. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16259–16268.

Zhao, Y.; Zhang, L.; Liu, Y.; Meng, D.; Cui, Z.; Gao, C.; Gao, X.; Lian, C.; and Shen, D. 2021b. Two-stream graph convolutional network for intra-oral scanner image segmentation. *IEEE Transactions on Medical Imaging*, 41(4): 826–835.

Zheng, Y.; Chen, B.; Shen, Y.; and Shen, K. 2022. TeethGNN: semantic 3D teeth segmentation with graph neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 29(7): 3158–3168.

Zou, B.-j.; Liu, S.-j.; Liao, S.-h.; Ding, X.; and Liang, Y. 2015. Interactive tooth partition of dental mesh base on tooth-target harmonic field. *Computers in biology and medicine*, 56: 132–144.