Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

VGDM: Visual Localization-Guided 3D Dental Segmentation via Extrinsic–Intrinsic Bridging

Hanbin Fang[1] , **Shidong Yang**[1] , **Fan Duan**[1]
Shuo Wang[2] , **YanHeng Zhou**[2] and **Li Chen**[1]
[1]Tsinghua University
[2]Beijing Tsinghua Changgung Hospital
fhb24@mails.tsinghua.edu.cn, stonehana1876@gmail.com, 18811520799@163.com
wangshuo_0724@sina.com, yanhengzhou@vip.163.com, chenlee@tsinghua.edu.cn

Abstract

3D dental segmentation is a key task in digital dentistry. In real intraoral scans data (IOS), occlusion, scanning noise, and reconstruction artifacts often break down the geometric separation structure between teeth, resulting in adjacent teeth being incorrectly merged or a single tooth being oversegmented. Since existing point cloud or mesh-based methods usually rely on local neighborhood consistency, when there are spurious geometric connections, features will diffuse across instances along geometric shortcuts, resulting in instance-level error propagation. To address this issue, we propose VGDM (Visual-Guided Diffusion Modulation), which serves as a bridge between extrinsic visual localization cues and intrinsic surface features under detection settings, enabling the extrinsic cues to regulate the propagation of intrinsic surface features. Instead of global propagation on the entire intraoral scan mesh, VGDM uses the single-view 2D detection results to roughly localize the tooth region and construct local 3D surface patches based on it. Within the patch, we soft-constrain feature propagation by visual cues to suppress the cross-instance propagation generated along spurious geometric connections, and introduce a dual-stream diffusion structure to improve the overall robustness. Experimental results on the largest public intraoral dataset(Teeth3DS) show that VGDM can significantly improve the segmentation rate of tooth instances and effectively reduce the merging and over-segmentation of adjacent teeth.

1 Introduction

3D dental crown segmentation is a fundamental task in digital dental diagnosis and treatment and plays a critical role in downstream applications such as 3D dental modeling, orthodontic design, and treatment planning [Zhang *et al.*, 2025]. The reliability of the segmentation results directly affects the stability of the subsequent process. Unfortunately, in real intraoral scans data(IOS), tooth occlusion, scanning noise, and reconstruction artifacts often destroy the geometric structure between teeth and introduce inconsistent connections with real tooth boundaries into the geometric representation.

In order to solve the automatic crown segmentation task, previous studies have proposed numerous point-based and mesh-based deep learning methods, gradually replacing the early traditional geometric methods [Cui *et al.*, 2021; Xiong *et al.*, 2023]. These methods usually model and propagate features based on local geometric neighborhoods, and have achieved good performance on public tooth segmentation datasets. However, under complex geometry conditions, even if the overall segmentation index is high, the segmentation results are still prone to structure-related errors.

In real dental scans, the restricted scanning angles and artifacts during the reconstruction process often disrupt the connection relationships between adjacent teeth, causing the originally anatomically separated teeth to have unreasonable connections. Current point cloud or mesh-based segmentation methods usually rely on these geometric connectivity relationships to transfer information between neighboring regions. Once the information is transmitted along the incorrect geometric connections, there is a tendency for information to be mixed between adjacent teeth, triggering structural errors such as merging or excessive segmentation.

Such structural errors caused by geometric connectivity are often difficult to naturally eliminate merely by enhancing the feature expression ability of the network. Even if the model can learn discriminative geometric features locally, when the information is transmitted along the incorrect geometric connections, its influence may still spread to adjacent regions, thereby interfering with the determination of instance boundaries. To mitigate this issue, one promising direction is to introduce robust *external* visual guidance. A rendered 2D view can provide relatively stable boundary cues even when the mesh surface contains unreasonable connections. Meanwhile, exploiting *intrinsic* feature consistency and geometric smoothness on the surface can regularize propagation under geometric noise and reduce boundary ambiguity.

Based on these observations, we propose *VGDM* (Visual-Guided Diffusion Modulation), a framework that enhances the robustness of feature propagation to geometry-induced spurious connections without explicitly modifying the connectivity of the 3D mesh. The core idea is a visually guided operator modulation strategy, which leverages relatively stable cues from a rendered 2D view to construct an

anisotropic modulation field on the 3D surface and modulate the Laplace–Beltrami operator (LBO) accordingly. With the modulated Laplace-Beltrami operator, we employ Laplace diffusion, since the LBO provides an intrinsic surface metric, yielding smoother and more coherent boundaries and making predictions less sensitive to local geometric noise. By increasing the resistance to feature diffusion across adhesion regions, VGDM effectively reduces over-segmentation of individual teeth and incorrect merging of adjacent teeth.

The main contributions are summarized as follows:

- We propose VGDM, a visual-guided diffusion framework for robust tooth instance segmentation under distorted geometric connectivity.
- We design an operator-level diffusion modulation mechanism guided by 2D visual information, which suppresses erroneous feature propagation across adhesion regions without altering mesh topology.
- We develop a dual-stream spectral diffusion framework that improves robustness to imperfect visual cues while preserving intrinsic geometric coherence, leading to more stable tooth instance segmentation.

2 Related Work

Accurate crown segmentation from IOS data is fundamental to digital dentistry. While many learning-based methods show promising results on standard public benchmarks [Sun *et al.*, 2020; Ben-Hamadou *et al.*, 2023], due to the scanning noise and reconstruction artifacts in the actual clinical IOS data [Fratila *et al.*, 2025], these factors can introduce unreasonable geometric connections on the mesh surface, which disrupt the originally clear anatomical boundaries between teeth. In such cases, tooth segmentation methods based on local geometric neighborhoods or fixed surface connection relationships are prone to structural errors such as incorrect fusion of adjacent teeth or excessive segmentation of single teeth. Most existing approaches operate on point clouds or meshes and aggregate local neighborhoods [Zanjani *et al.*, 2019; Xiong *et al.*, 2023; Duan and Chen, 2023], sometimes with task-specific post-processing or multi-stage pipelines. Their core assumption is that local Euclidean neighborhoods or fixed surface connectivity reliably reflect tooth boundaries. In challenging IOS (adhesion, noise, partial scans), this assumption breaks and can cause cross-tooth mis-merging or over-segmentation.

Most existing point cloud or mesh-based segmentation methods rely on the geometric neighborhood relationships for feature modeling and information transmission in the spatial domain. The early PointNet series of methods introduced modeling of local geometric structures through local region division and feature aggregation [Qi *et al.*, 2017a; Qi *et al.*, 2017b]; subsequently, methods based on graph neural networks explicitly constructed neighborhood graphs and performed message passing based on geometric connectivity, thereby enhancing the local context modeling ability [Wang *et al.*, 2019]. And spectral methods such as DiffusionNet [Sharp *et al.*, 2022] have further modeled feature propagation as a diffusion process defined on surfaces, achieving modeling of the intrinsic geometric structure through the Laplacian-Beltrami operator [Reuter *et al.*, 2009], demonstrating advantages in maintaining coordinate invariance and resolution stability. Although these methods model the geometric structure at different levels, their common feature is that the feature transmission process is driven by local geometric neighborhoods or surface connectivity, and implicitly assumes that these geometric relationships can reliably reflect the true anatomical boundaries. When artifacts or adhesion phenomena introduce incorrect geometric connections, this type of feature transmission based on geometric connectivity may occur along unreasonable paths, resulting in cross-instance information mixing and structural segmentation errors.

3 Method

3.1 Problem Formulation

We represent each IOS as a triangular mesh surface $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ denotes the vertex set with vertex coordinates $v_i \in \mathbb{R}^3$, and $\mathcal{E}$ and $\mathcal{F}$ denote the sets of mesh edges and triangular faces, respectively. For each vertex v_i, we construct an input feature vector $x_i \in \mathbb{R}^d$, such as coordinates, Gaussian curvatures, or other local geometric descriptors, and stack them as

$$X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times d}.$$

Our goal is tooth-wise instance segmentation: given $\mathcal{M}$ and X, we predict a per-vertex instance assignment $\hat{z}_i \in \{1, \ldots, \hat{K}\}$ and a label predict $\mathcal{L}$, where K denotes the (ground-truth) number of tooth instances in the scan (varying across samples) and $\hat{K}$ denotes the number of predicted instances, and $\mathcal{L}$ represents the prediction from the predicted instance to the real tooth FDI label. Equivalently, the model may output per-vertex embeddings $e_i \in \mathbb{R}^q$ and obtain instances via clustering or connected-component grouping. We denote the final set of predicted instances as $\mathcal{P} = \{p_j\}_{j=1}^{\hat{K}}$, where each $p_j \subseteq \mathcal{V}$ is the vertex subset corresponding to a predicted tooth instance.

3.2 Overview

Given the IOS mesh $\mathcal{M}$ and its vertex features X, we propose *Visual-Guided Diffusion Modulation* (VGDM) for tooth instance segmentation. The core idea of VGDM is to use single-view visual cues to constrain information propagation on the mesh - increasing the cross-edge propagation impedance in potential inter-tooth regions, making it more difficult for features to diffuse across tooth gaps, thereby suppressing cross-tooth false coupling and reducing instance-level merge/split errors.

As shown in Figure 1, the overall process of VGDM is summarized as follows:

(1) Visual Cues Modeling. We render the mesh from a near-occlusal view and apply a 2D vision model to obtain dense toothness predictions. These cues are then projected back onto the mesh to produce vertex-wise visual confidence, and the 2D detections are used to localize candidate tooth regions as local 3D surface patches.

(2) Diffusion Modulation. Based on the projected cues, we construct two Laplace–Beltrami operators on the same

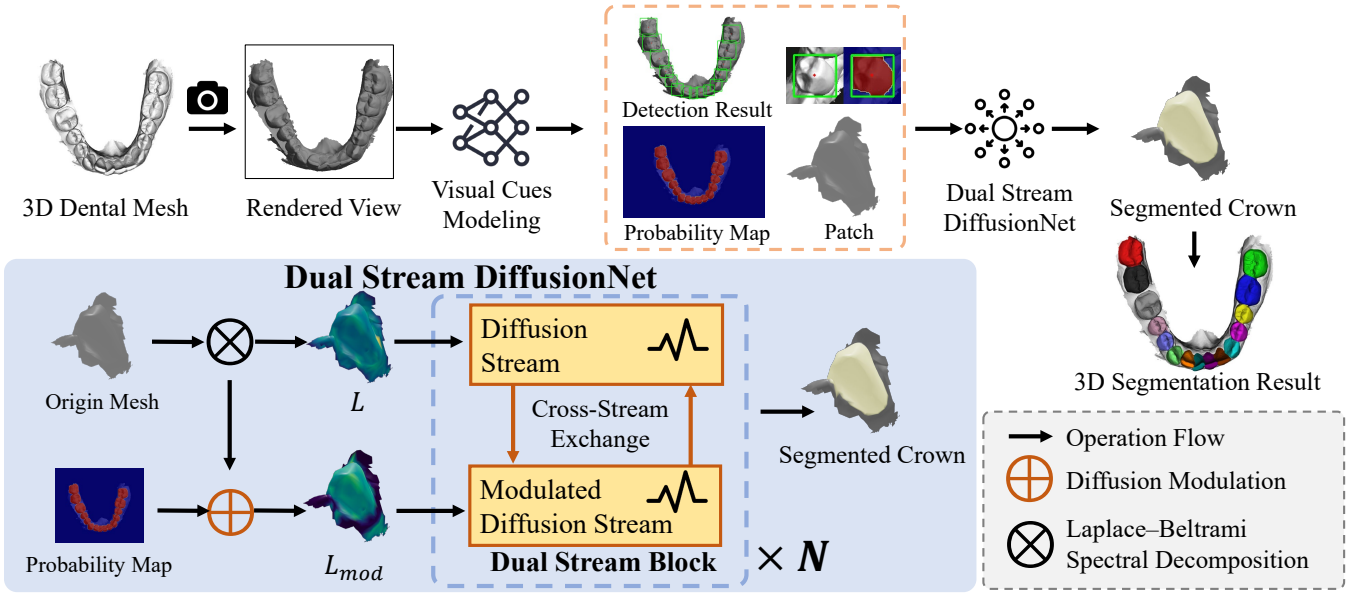

Figure 1: Overview of the proposed VGDM. Visual cues from a single rendered view are projected onto the mesh to modulate the Laplace–Beltrami operator, enabling a dual-stream diffusion architecture that suppresses cross-tooth feature leakage caused by spurious geometric connections.

patch: the original operator L and a visually modulated operator L_{mod}. While L preserves intrinsic geometric diffusion, L_{mod} down-weights diffusion conductance across potential inter-tooth boundaries, preventing feature leakage to adjacent teeth.

(3) Dual-stream DiffusionNet. We adopt DiffusionNet as the 3D backbone (where "diffusion" refers to intrinsic PDE-based feature propagation rather than generative diffusion). VGDM employs a dual-stream architecture that propagates features under L and L_{mod} in parallel, and fuses the two streams to produce final predictions for each patch. This design improves robustness to adhesion-induced structural errors while mitigating imperfect 2D visual cues.

3.3 Visual Cues Modeling

This section describes how we obtain single-view visual cues and visibility, and map them back to the mesh. We render the mesh from a near-occlusal view (an approximate top-down direction aligned with the occlusal-plane normal) to obtain a 2D representation, and apply a plug-and-play 2D detection model to predict a per-pixel *toothness* probability map. In our implementation, we use a YOLO-family model, but this component is not tied to any specific architecture and can be replaced by any 2D network that outputs toothness probabilities/confidences. The near-occlusal rendering is used to obtain dense 2D toothness predictions, which are mapped back to the mesh to form a per-vertex visual probability field P and a corresponding visibility mask m, serving as inputs to subsequent operator modulation and dual-stream diffusion. Each 2D tooth proposal is then back-projected onto the 3D mesh to define a local surface patch corresponding to a candidate tooth region. Formally, for a 2D proposal box b_t, we define the corresponding patch as $\mathcal{P}_t = \{v_i \in V \mid \pi(v_i) \in \text{expand}(b_t)\}$, where $\pi(\cdot)$ denotes the projection and $\text{expand}(\cdot)$ slightly enlarges the box to include boundary regions.In this way, the 2D boxes induce a set of patch-level regions on the mesh, which serve as the basic units for subsequent visual guidance and diffusion modulation. At this stage, the 2D model also predicts an FDI label $\mathcal{L}$ for each proposal, which is used to specify the tooth position of the corresponding 3D patch.

Toothness Prediction and Visibility. During rendering, we also compute per-vertex visibility (e.g., via a z-buffer visibility mask) to indicate whether each vertex is observable from the chosen view. Let $P_{2\text{D}}$ denote the predicted 2D toothness probability map and $\pi(v_i)$ the projected pixel coordinate of vertex v_i. We obtain per-vertex toothness probabilities by bilinear sampling: $P_i = \text{bilinear}(P_{2\text{D}}, \pi(v_i))$. We additionally obtain a per-vertex visibility mask $m_i \in \{0, 1\}$, where $m = 1$ for visible vertices and $m = 0$ for invisible ones. The resulting P and m are used as inputs to diffusion modulation: P provides visual confidence variations for constructing diffusion barriers, while m suppresses unreliable modulation in occluded or non-visible regions.

3.4 Diffusion Modulation

This section describes how we construct a pair of spectral operators on the same mesh topology by reweighting surface conductance using single-view visual cues and visibility. The key motivation is that real IOS meshes may contain spurious short-circuit connections (e.g., tooth adhesion), under which spectral diffusion methods—including DiffusionNet—tend to propagate features across incorrect topology, leading to merge/split failures. Instead of explicitly modifying the mesh topology, we intervene before diffusion by modulating the conductance of the Laplacian operator. The goal

is to construct a visually modulated Laplacian L_{mod} such that diffusion within a tooth is largely preserved, diffusion across tooth boundaries is strongly suppressed, and the behavior remains consistent under mesh refinement.

Spectral Method. In surface and mesh learning, *spectral method* refers to approaches based on the Laplace–Beltrami operator (and its discretization) to characterize intrinsic geometry. The eigenfunctions and eigenvalues of the Laplacian (the *spectrum*) provide a multi-scale basis for analyzing functions on the surface. Intuitively, the Laplacian governs smoothing and diffusion: low-frequency components capture global structure, while higher frequencies encode local details and boundaries. Spectral diffusion and filtering enable information propagation that depends weakly on the specific mesh discretization, which underlies their relative stability across different resolutions.

Spectral Preliminaries. We use the discrete Laplace–Beltrami operator L together with a mass matrix M to model intrinsic surface diffusion, and compute the generalized eigendecomposition $L\phi_k = \lambda_k M\phi_k$ to obtain the first R eigenpairs (λ_k, ϕ_k). These eigenpairs are used to construct the spectral operators underlying DiffusionNet and to compute spectral input features such as the heat kernel signature (HKS), which characterizes how heat diffuses locally around each vertex over multiple scales, providing an intrinsic and robust description of local surface geometry. The resulting HKS is included as part of the per-vertex input feature vector to the network. In what follows, we further construct a visually modulated operator L_{mod} and its corresponding spectrum $(\lambda_k^{mod}, \phi_k^{mod})$ to form a dual-operator setting.

Original Operator. We adopt the standard cotangent discretization of the Laplace–Beltrami operator. For each edge $(i, j) \in \mathcal{E}$, the undirected cotangent weight is

$$w_{ij}^{\mathrm{orig}} = \frac{1}{2}\big(\cot \alpha_{ij} + \cot \beta_{ij}\big), \tag{1}$$

where α_{ij} and β_{ij} are the two angles opposite to edge (i, j) (for boundary edges, only one angle is available). This yields a symmetric stiffness matrix L with off-diagonal entries $L_{ij} = -w_{ij}^{\mathrm{orig}}$, together with a lumped (or barycentric) mass matrix M for the generalized eigendecomposition.

Diffusion Modulation. Let $e_{ij} = v_i - v_j$ be the edge vector and using P_i as defined above. We introduce a small constant $\delta > 0$ for numerical stability. Our design principle is to modify the *conductance of the diffusion medium* prior to constructing the Laplacian. For each edge, we define a *refinement-consistent* barrier energy $q_{ij} = \frac{(P_i - P_j)^2}{\|e_{ij}\| + \delta}$. This form can be viewed as a discrete counterpart of the continuous energy $\int (P'(s))^2\, ds$, and behaves consistently under mesh refinement: when an edge is subdivided, the accumulated barrier over the sub-edges remains approximately preserved. We then convert this barrier energy into a conductance scaling factor using an exponential model:

$$s_{ij} = \begin{cases} 1, & \text{if } (m_i = 0) \text{ or } (m_j = 0), \\ \varepsilon + (1-\varepsilon) G_c e^{-\alpha q_{ij}}, & \text{if } m_i = m_j = 1, \end{cases} \tag{2}$$

where m_i and m_j means If there is an endpoint that is not visible, the scaling factor will always be 1, and $G_c = \sqrt{P_i P_j}$ acts as a confidence gate that globally downweights low-confidence regions for robustness, and $\exp(-\alpha q_{ij})$ enforces exponential suppression of diffusion across visual boundaries (large $|P_i - P_j|$). The small constant ε prevents numerical degeneracy and is not intended to repair connectivity. Note that this exponential term is fundamentally different from the heat-kernel diffusion $\exp(-tL)$: here it defines the diffusion *medium* before propagation, whereas the latter performs diffusion *within* a given medium.

The modulated edge weights are then obtained by relative reweighting:$w_{ij}^{\mathrm{mod}} = w_{ij}^{\mathrm{orig}} \cdot s_{ij}$, and the visually modulated Laplacian is reconstructed as $L_{\mathrm{mod}} = D_{\mathrm{mod}} - W_{\mathrm{mod}}$.

This construction preserves symmetry and non-negativity ($w_{ij}^{\mathrm{mod}} = w_{ji}^{\mathrm{mod}} \geq 0$), ensuring that L_{mod} remains a valid Laplacian-type operator. After get L_{mod},only need to fix M, we can get $(\lambda_k^{mod}, \phi_k^{mod})$ from $L_{\mathrm{mod}}\phi_k^{mod} = \lambda_k^{mod} M\phi_k^{mod}$.

3.5 Dual-stream DiffusionNet

We adopt DiffusionNet [Sharp *et al.*, 2022] as the 3D backbone for feature propagation (here, "diffusion" refers to PDE-driven *geometric* diffusion on surfaces, *not* generative diffusion models).Using DiffusionNet as a base module, we build a dual-stream architecture comprising an original spectral decomposition and a visually modulated spectral decomposition. For each sample, the dataloader provides two sets of spectral decompositions: (λ_k, ϕ_k) and $(\lambda_k^{mod}, \phi_k^{mod})$; both are fed into the network during training to enable parallel propagation and interaction.

Dual-stream Backbone. Given per-vertex input features x_{in}, we first map them to a common channel width C using a shared linear layer:

$$x^0 = \mathrm{Lin}_{\mathrm{first}}(x_{\mathrm{in}}). \tag{3}$$

We duplicate the result to initialize the two streams. For the t-th block ($t = 0, \ldots, N_{\mathrm{block}} - 1$), the two streams are updated by their respective DiffusionNet blocks parameterized by different eigenpairs: ($x_{\mathrm{orig}}^{t+1} = \mathrm{Block}_{\mathrm{orig}}^t\big(x_{\mathrm{orig}}^t;\, L\big)$, $x_{\mathrm{mod}}^{t+1} = \mathrm{Block}_{\mathrm{mod}}^t(x_{\mathrm{mod}}^t;\, L_{\mathrm{mod}})$). The two streams use independent DiffusionNet block parameters while operating on the original and visually modulated Laplacian operators, respectively.

DiffusionNet Block. A DiffusionNetBlock performs intrinsic feature mixing on the vertex set using diffusion/filtering operators parameterized by the Laplace–Beltrami operator L. It is a *geometric* diffusion layer (not a generative diffusion model): the propagation is driven by surface operators, remains stable across mesh resolutions and sampling densities, and naturally supports variable-length vertex inputs.

Block-wise Cross-stream Exchange. To exploit the complementarity between the two propagation structures, we introduce a cross-stream exchange module after each block. Specifically, we form a difference-enhanced feature $f^{i+1} = \mathrm{concat}\Big(x_{\mathrm{orig}}^{i+1},\, x_{\mathrm{mod}}^{i+1},\, x_{\mathrm{orig}}^{i+1} - x_{\mathrm{mod}}^{i+1}\Big)$, and use two lightweight mappings $\mathrm{Exch}_{\text{to-orig}}^t$ and $\mathrm{Exch}_{\text{to-mod}}^t$ to generate exchange signals injected into the two streams. We further

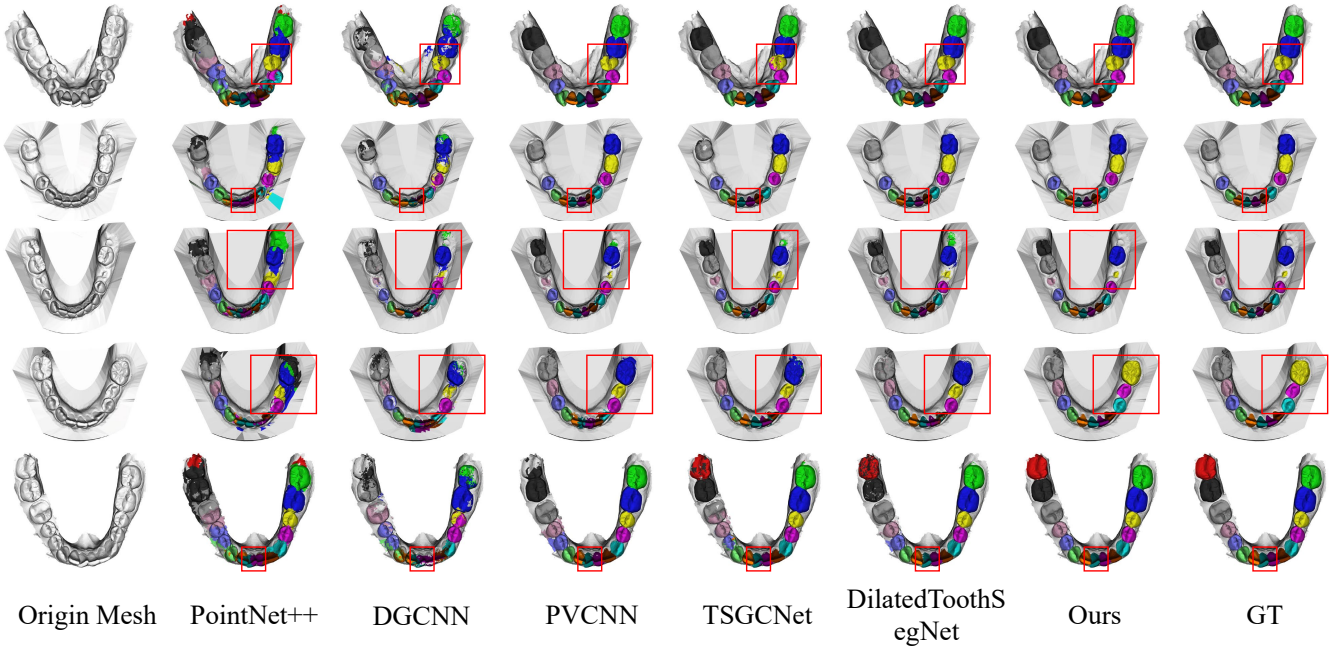

Figure 2: Qualitative comparison of dental segmentation results from six models on five challenging cases, including crowded dentition, missing teeth, and the presence of wisdom teeth, as well as scans with noise. Different colors indicate tooth instance labels. Our method preserves cleaner boundaries and remains robust to noise, whereas other methods often introduce boundary artifacts.

introduce learnable per-channel gating vectors $\alpha^t_{\text{orig}}, \alpha^t_{\text{mod}} \in \mathbb{R}^C$ to control the exchange strength:

$$\begin{aligned} x^{i+1}_{\text{orig}} &\leftarrow x^{i+1}_{\text{orig}} + \sigma(\alpha^i_{\text{orig}}) \odot \text{Exch}^i_{\text{to-orig}}(f^{i+1}), \\ x^{i+1}_{\text{mod}} &\leftarrow x^{i+1}_{\text{mod}} + \sigma(\alpha^i_{\text{mod}}) \odot \text{Exch}^i_{\text{to-mod}}(f^{i+1}), \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ is channel-wise multiplication. This design allows the network to learn *when* and *which channels* to exchange information, mitigating potential error amplification when the modulated stream is unreliable.

Robust Late Fusion and Output. At the final stage of the network, we adopt a robust late fusion strategy where the original stream serves as the baseline, and the modulated stream provides residual corrections. We denote the final outputs of the two streams as $x_{\text{orig}} = x^{N_{\text{block}}}_{\text{orig}}$ and $x_{\text{mod}} = x^{N_{\text{block}}}_{\text{mod}}$. Let $f_{\text{fuse}} = \text{concat}(x_{\text{orig}}, x_{\text{mod}}, |x_{\text{orig}} - x_{\text{mod}}|)$ The fused feature is computed as $x = x_{\text{orig}} + \text{Lin}_{\text{fuse}}(f_{\text{fuse}})$ and through a linear output head and a sigmoid function, each node's probability is output.we output a patch-specific foreground probability $p_t(v_i)$ (tooth vs. non-tooth) for each patch

On each patch, we extract the largest connected foreground as our tooth segmentation result. For the patch-to-mesh assembly, we employ a tooth-centric instance resolution strategy. Specifically, we group patches by their assigned FDI labels and calculate a confidence score based on the sum of predicted foreground probabilities. For each tooth label, the system automatically selects the patch with the highest score as the optimal candidate, effectively resolving overlap conflicts without complex global optimization.

4 Experiments

4.1 Experiment Setup

We conducted experiments on the Teeth3DS dataset [Ben-Hamadou *et al.*, 2023], which contains 900 upper tooth scan samples and 900 lower tooth scan samples. This dataset provides per-vertex annotations of tooth instances, with the tooth instances labeled using the FDI numbering system to distinguish individual permanent teeth within different quadrants. This numbering system is widely used in clinical dentistry to uniquely identify permanent teeth at different positions. The dataset also includes normal samples as well as samples with complex conditions such as wisdom teeth, tooth loss, or tooth fusion, to assess the model's performance in tooth instance segmentation in scenarios close to real clinical settings.

4.2 Implementation Details

All experiments are implemented in PyTorch. We adopt DiffusionNet as the base network architecture to perform feature propagation and prediction on the mesh surface. The network input includes the 3D coordinates of vertices, Gaussian curvature, Heat Kernel Signature (HKS), and the tooth probability map from the 2D detection model. We use yolo11m-seg as the detector, which is pre-trained before 3D training.

In the dual-stream design, the two branches share the same architecture and training configuration, but perform diffusion propagation with two different operators, L and L_{mod}, respectively, followed by feature exchange and fusion. The model is trained with Adam using an initial learning rate of 5×10^{-4} for 200 epochs. During the inference stage, we threshold per-vertex probabilities and extract the largest connected component to form tooth instances. To ensure fair-

Method	T_{mIOU}	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]	67.65	77.58	71.63	70.90	68.20	71.67	66.98	67.66	55.65	-
TSegNet [Cui *et al.*, 2021]	59.81	67.35	57.05	62.89	69.03	58.74	52.23	61.45	66.02	-
TeethGNN [Zheng *et al.*, 2022]	83.89	88.46	86.58	86.31	86.56	87.69	82.05	79.21	82.89	66.54
TSRNet [Jin *et al.*, 2024]	86.56	90.91	88.20	87.83	87.35	87.65	86.57	86.69	83.74	70.82
ToothGroupNet [Ben-Hamadou *et al.*, 2023]	90.16	92.88	92.19	92.30	92.65	93.44	87.74	88.84	84.82	68.20
TSGCNet [Zhang *et al.*, 2021]	79.79	86.73	82.63	81.59	82.82	82.75	80.04	80.60	69.27	34.23
IsbNet [Ngo *et al.*, 2023]	81.01	87.65	68.61	80.62	84.06	86.44	84.61	85.79	80.44	26.75
DGCNN [Wang *et al.*, 2019]	80.08	87.27	80.85	80.31	81.12	82.67	79.86	81.53	75.45	52.42
CBAnet [Jin *et al.*, 2025]	86.77	91.16	88.62	87.94	88.17	88.41	86.39	86.12	83.05	76.88
PT [Zhao *et al.*, 2021]	71.52	78.84	71.26	72.15	72.08	72.50	69.55	73.27	74.88	2.27
DilatedSegNet [Krenmayr *et al.*, 2024]	86.55	91.26	86.49	86.46	87.33	88.61	86.83	88.09	83.88	62.49
3DTeethSAM [Lu *et al.*, 2025]	91.90	94.33	93.28	93.36	93.16	93.79	91.65	90.73	89.10	83.29
Ours	91.20	94.25	93.13	93.54	93.48	93.83	90.53	90.16	90.47	84.29
Ours*	94.98	97.27	94.39	94.64	95.13	96.11	96.22	96.79	94.36	95.72

Table 1: We adopt the official data division method from the MICCAI challenge, which used 1200 scanned samples as the training set and 600 scanned samples as the test set. T_{mIOU} is the average for the mean IoU over all tooth classes rather than calculating the average for each sample, respectively. Columns T1/9 through T8/16 present the per-tooth mIoU scores, with labels grouped for symmetric positions.

ness in comparison, all comparison methods use the same post-processing procedure. The parameter of our model is 3.67M. Regarding computational cost, while the pipeline includes 2D detection and patch-wise spectral decomposition, the overhead is well-managed via parallelization. Given that dental CAD workflows are mostly offline, this minor latency is a highly worthwhile trade-off for the significant reduction in structural errors.

4.3 Evaluation Metrics

For more clearly define the instance-level errors of teeth, we define the instance level GT set $\mathcal{G} = \{G_i\}_{i=1}^{K}$ and the predicted instance set $\mathcal{P} = \{p_j\}_{j=1}^{\hat{K}}$, where G_i and p_j represent the set of vertices corresponding to the i-th real tooth instance and the j-th predicted instance (ignoring the gum vertices), respectively. We adopt the instance-level intersection and union ratio $IoU(G_i, p_j) = \frac{|G_i \cap p_j|}{|G_i \cup p_j|}$, And for each real instance, we take the IoU value of its best matching predicted instance $best\_iou(i) = \max_j IoU(G_i, p_j)$. Given the threshold τ, the TSR is defined as

$$TSR@\tau = \frac{1}{K} \sum_{i=1}^{K} \left[\text{best\_iou}(i) \geq \tau\right], \tag{5}$$

where $[\cdot]$ represents the indicator function, which takes the value 1 if the condition in parentheses holds and 0 otherwise. Secondly, in order to characterize structural errors such as merging and fragmentation, we use the asymmetric coverage ratio:$\text{Cov}(\text{A}, \text{B}) = \frac{|A \cap B|}{|A|}$. We can also define the Merge rate to count the cases where multiple true teeth were incorrectly merged into the same predicted instance. We first define the main match as $\phi(i) = \arg\max_j \text{Cov}(G_i, p_j)$, then the overall merge rate is defined as follows.

$$\text{Merge@}\tau = \frac{1}{K} \sum_{i=1}^{K} \left[\exists\, i' \neq i :\ \text{Cov}(p_{\phi(i)}, G_{i'}) \geq \tau\right], \tag{6}$$

Similarly, we define the Split rate to count the number of instances where a single real tooth is over-segmented into multiple predicted instances:

$$\text{Split@}\tau = \frac{1}{K} \sum_{i=1}^{K} \Big[\exists\, j \neq k :\ \text{Cov}(G_i, p_j) \geq \tau \wedge \text{Cov}(G_i, p_k) \geq \tau\Big]. \tag{7}$$

In order to facilitate the comparison with the existing work, we use the commonly used evaluation metrics in the literature alignment experiment setting, and report the results according to their data division and evaluation process.

4.4 Main Results

As shown in Table 1, our full pipeline ("ours") achieves competitive performance across all metrics, reaching a T_{mIOU} of 91.20% and a Dice score of 94.25%, which is on par with or superior to most existing methods. Notably, our approach demonstrates stable and balanced performance across different tooth indices, indicating strong robustness of the proposed 3D diffusion-based modulation and dual-stream propagation under realistic conditions. To further analyze the upper limit of our method, we report an "our*" optimized version, in which the errors introduced in the 2D inspection stage (i.e., incorrect label predictions and missed bounding boxes) have been excluded. This setting is used to simulate the maximum potential capability of our 3D framework under the condition of having perfect 2D prior conditions. Under this idealized condition, our method achieves significant improvements, with the mIOU value increasing from 91.20% to 94.98%, and the Dice value increasing from 94.25% to 97.27%. These improvements are particularly significant in the posterior teeth area, as severe occlusal problems and adhesions between teeth are more common here, and in fact, the main reason for the current limitation of posterior teeth performance lies in the 2D visual module. These results indicate that the performance bottleneck of the entire process mainly stems from imperfect 2D predictions, while the proposed 3D diffusion modulation framework itself has a relatively high upper limit of performance.

Method	TSR@0.5↑	TSR@0.75↑	Merge@0.1↓	Merge@0.2↓	Split@0.1↓	Split@0.2↓
PointNet++ [Qi *et al.*, 2017b]	58.41	11.43	45.08	26.61	42.44	22.58
DGCNN [Wang *et al.*, 2019]	59.95	32.18	35.40	25.59	28.25	15.33
PVCNN [Liu *et al.*, 2019]	72.68	68.22	5.90	3.28	5.34	2.64
TSGCNet [Zhang *et al.*, 2021]	97.51	90.90	3.43	1.99	3.34	1.56
DilatedToothSegNet [Krenmayr *et al.*, 2024]	<u>99.27</u>	<u>97.01</u>	<u>1.37</u>	<u>0.91</u>	<u>1.10</u>	<u>0.50</u>
Ours	**99.70**	**99.18**	**0.25**	**0.21**	**0.06**	**0.06**

Table 2: TSR indicates whether teeth are correctly separated, and merge and split rates quantify instance-level merge and oversegmentation errors, which we compute at different coverage thresholds and report simultaneously. Higher TSR and lower Merge/Split indicate better performance. The bold and underlined values indicate the best and second-best results, respectively.

4.5 Split & Merge Study

While many existing methods report high segmentation accuracy, they still suffer from severe structural errors under challenging intraoral scan conditions, especially tooth adhesion and occlusal noise. Such errors mainly manifest as merging adjacent teeth or over-segmenting a single tooth, which are not adequately reflected by conventional vertex-level metrics. Therefore, we conduct a dedicated split-and-merge study to evaluate the structural correctness of tooth instance segmentation. Table 2 reports the instance-level performance of different methods on the Teeth3DS dataset using TSR, Merge, and Split metrics under multiple thresholds. Generic point-based methods such as PointNet++ and DGCNN exhibit low TSR values and high merge/split rates, indicating their limited ability to handle distorted local neighborhoods caused by interdental adhesion. Methods specifically designed for dental segmentation (e.g., TSGCNet and DilatedToothSegNet) achieve higher TSR, demonstrating improved tooth separation. However, they still suffer from non-negligible merge and split errors, especially under stricter coverage thresholds, suggesting that incorrect feature propagation across adhesive regions remains an issue. In contrast, our dual-stream VGDM achieves the highest TSR scores (99.70% @0.5 and 99.18% @0.75) while maintaining near-zero merge and split rates across all thresholds. This demonstrates that VGDM is not only capable of correctly separating adjacent teeth but also highly effective in suppressing both false merging and over-segmentation.

Qualitative results in Figure 2 further support this observation. While baseline methods frequently produce merged crowns or fragmented tooth instances in adhesive regions, our method consistently yields clean and anatomically plausible tooth boundaries, closely matching the ground truth. This confirms that the proposed operator-level modulation successfully mitigates diffusion leakage across spurious geometric connections. Moreover, using Laplace–Beltrami diffusion as an intrinsic low-pass regularizer promotes smooth and coherent boundaries, making the predictions less sensitive to local geometric noise.

4.6 Ablation Study

The instance segmentation results of different model variants on the Teeth3DS dataset are shown in Table 3. Firstly, by comparing the full model with the L model, although the mIoU differences among variants are small, it can be found that the dual-stream model is significantly better than the single-stream model in terms of Merge and Split indicators, indicating that the parallel diffusion based on different operators is helpful to alleviate the structural confusion between adjacent teeth. Secondly, by comparing the full model with L_{mod} model, it can be found that the model cannot give good segmentation results due to the lack of real geometry input.

Setting	mIOU↑	TSR↑	Merge↓	Split↓
Full Model	**95.21**	**95.89**	**16.22**	**15.55**
w/o Metric Surgery	95.11	95.16	20.43	19.89
w/o Dual-Stream	67.63	1.13	95.54	97.34

Table 3: Ablation study on core architectural components. Removing Metric Surgery causes failures on adhesive teeth (low TSR). Removing Dual-Stream increases structural errors and may cause failure cases when visual cues are unreliable. We select TSR@0.90, Merge@0.01, and Split@0.01 for this table.

Detector	mAP_{50}↑	Precision↑	Recall↑	F1-Score↑
Mask	96.92	94.71	94.41	94.55
box	96.90	94.68	94.38	94.53

Table 4: Performance of the 2D detector for tooth localization.

Furthermore, Table 4 reports performance of the adopted YOLOv11 model under two output forms. The results show that even under relatively general 2D segmentation results, the proposed VGDM can still work stably and show better segmentation results.

5 Conclusion

We introduce VGDM, a novel 3D tooth segmentation framework. Our method has achieved competitive results in the Teeth3DS benchmark test, especially in reducing instance-level merging and segmentation errors, which are crucial for reliable clinical applications. The proposed framework demonstrates strong robustness under complex anatomical conditions, indicating its suitability for real-world digital dental workflow. More broadly speaking, VGDM highlights the potential of integrating external visual priors and internal geometric operators at the diffusion level. We believe that this paradigm can go beyond tooth segmentation and be extended to other surface-based three-dimensional perception tasks, where false topologies and ambiguous geometries pose fundamental challenges.

Acknowledgements

We are grateful for the valuable comments provided by the anonymous reviewers. This research was partially supported by the National Natural Science Foundation of China (Grant No. 61972221).

References

[Ben-Hamadou *et al.*, 2023] Achraf Ben-Hamadou, Oussama Smaoui, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Hoyeon Lim, Minchang Kim, Minkyung Lee, Minyoung Chung, Yeong-Gil Shin, et al. 3dteethseg'22: 3d teeth scan segmentation and labeling challenge. *arXiv preprint arXiv:2305.18277*, 2023.

[Cui *et al.*, 2021] Zhiming Cui, Changjian Li, Nenglun Chen, Guodong Wei, Runnan Chen, Yuanfeng Zhou, Dinggang Shen, and Wenping Wang. Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model. *Medical Image Analysis*, 69:101949, 2021.

[Duan and Chen, 2023] Fan Duan and Li Chen. 3d dental mesh segmentation using semantics-based feature learning with graph-transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 456–465. Springer, 2023.

[Fratila *et al.*, 2025] Anca Maria Fratila, Adriana Saceleanu, Vasile Calin Arcas, Nicu Fratila, and Kamel Earar. Enhancing intraoral scanning accuracy: From the influencing factors to a procedural guideline. *Journal of Clinical Medicine*, 14(10):3562, 2025.

[Jin *et al.*, 2024] Hairong Jin, Yuefan Shen, Jianwen Lou, Kun Zhou, and Youyi Zheng. Tsrnet: A dual-stream network for refining 3d tooth segmentation. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[Jin *et al.*, 2025] Hairong Jin, Jianwen Lou, Zhiguo Lu, Teng Wu, Kun Zhou, and Youyi Zheng. Learning center-and boundary-aware instance representation for 3d tooth segmentation. *Computers & Graphics*, page 104313, 2025.

[Krenmayr *et al.*, 2024] Lucas Krenmayr, Reinhold von Schwerin, Daniel Schaudt, Pascal Riedel, and Alexander Hafner. Dilatedtoothsegnet: Tooth segmentation network on 3d dental meshes through increasing receptive vision. *Journal of Imaging Informatics in Medicine*, 37(4):1846–1862, 2024.

[Liu *et al.*, 2019] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in neural information processing systems*, 32, 2019.

[Lu *et al.*, 2025] Zhiguo Lu, Jianwen Lou, Mingjun Ma, Hairong Jin, Youyi Zheng, and Kun Zhou. 3dteethsam: Taming sam2 for 3d teeth segmentation. *arXiv preprint arXiv:2512.11557*, 2025.

[Ngo *et al.*, 2023] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. 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*, pages 13550–13559, 2023.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[Reuter *et al.*, 2009] Martin Reuter, Silvia Biasotti, Daniela Giorgi, Giuseppe Patanè, and Michela Spagnuolo. Discrete laplace–beltrami operators for shape analysis and segmentation. *Computers & Graphics*, 33(3):381–390, 2009.

[Sharp *et al.*, 2022] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022.

[Sun *et al.*, 2020] Diya Sun, Yuru Pei, Peixin Li, Guangying Song, Yuke Guo, Hongbin Zha, and Tianmin Xu. Automatic tooth segmentation and dense correspondence of 3d dental model. In *International conference on medical image computing and computer-assisted intervention*, pages 703–712. Springer, 2020.

[Wang *et al.*, 2019] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[Wu *et al.*, 2022] Tai-Hsien Wu, Chunfeng Lian, Sanghee Lee, Matthew Pastewait, Christian Piers, Jie Liu, Fan Wang, Li Wang, Chiung-Ying Chiu, Wenchi Wang, et al. 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, 2022.

[Xiong *et al.*, 2023] Huimin Xiong, Kunle Li, Kaiyuan Tan, Yang Feng, Joey Tianyi Zhou, Jin Hao, Haochao Ying, Jian Wu, and Zuozhu Liu. Tsegformer: 3d tooth segmentation in intraoral scans with geometry guided transformer. In *International conference on medical image computing and computer-assisted intervention*, pages 421–432. Springer, 2023.

[Zanjani *et al.*, 2019] Farhad Ghazvinian Zanjani, David Anssari Moin, Bas Verheij, Frank Claessen, Teo Cherici, Tao Tan, et al. Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In *International Conference on Medical Imaging with Deep Learning*, pages 557–571. PMLR, 2019.

[Zhang *et al.*, 2021] Lingming Zhang, Yue Zhao, Deyu Meng, Zhiming Cui, Chenqiang Gao, Xinbo Gao, Chunfeng Lian, and Dinggang Shen. 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 (CVPR)*, pages 6699–6708, June 2021.

[Zhang *et al.*, 2025] Haotian Zhang, Tianran Yuan, Tingcheng Li, Juan Du, Maokang Ye, and Qian Jiang. A review of tooth ai segmentation on medical data. *Discover Imaging*, 2(1):17, 2025.

[Zhao *et al.*, 2021] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[Zheng *et al.*, 2022] Youyi Zheng, Beijia Chen, Yuefan Shen, and Kaidi Shen. Teethgnn: semantic 3d teeth segmentation with graph neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 29(7):3158–3168, 2022.