ARCH-Net: Aligned 3D Teeth Reconstruction via Counterfactual-Enhanced Hierarchical Framework

Qingxin Deng[†1], Landu Jiang[†2], Yuzheng Yan[1], Wuman Luo[*1], and Dian Zhang[*3]

[1] Macao Polytechnic University, Rua de Luís Gonzaga Gomes, Macao SAR, China luowuman@mpu.edu.mo

[2] The Hongkong University of Science and Technology (Guangzhou), Duxue Road 1, Guangzhou, China landujiang@hkust-gz.edu.cn

[3] Lingnan University, 8 Castle Peak Road, Tuen Mun Hong Kong SAR, China zhangdian@LN.edu.hk

Abstract. Accurate 3D dental reconstruction from sparse, uncalibrated intra-oral photographs is critical for radiation-free digital dentistry. However, traditional parametric methods lack strong anatomical priors, are restricted to limited parameter spaces, and exhibit high inference latency. Meanwhile recent generative approaches often lack geometric rigor and cross-view consistency. We propose ARCH-Net, an Aligned 3D teeth Reconstruction approach via a Counterfactual-enhanced Hierarchical framework. ARCH-Net systematically decomposes the task into three progressive stages: (1) Anatomically-constrained latent space construction to capture robust geometric priors within a learned embedding space; (2) Counterfactual-enhanced 2D-3D alignment, where a latent diffusion model is supervised by a novel counterfactual learning strategy. By contrasting predicted latent codes against mismatched negative samples, this strategy significantly amplifies the model's discriminative power to enforce rigorous geometric consistency across sparse views; and (3) Prior-driven mesh refinement via sequential rigid and non-rigid registration to ensure the topological integrity of final high-resolution models. Experimental results demonstrate that ARCH-Net significantly outperforms state-of-the-art methods in both reconstruction accuracy and computational efficiency. Furthermore, our approach is robust to varying imaging conditions, offering a reliable, radiation-free solution for precision orthodontic planning and real-time digital diagnostics.

Keywords: 3D teeth reconstruction · 2D-3D alignment · Diffusion model.

1 Introduction

Orthodontic treatment plays a fundamental role in modern dentistry, focusing on the restoration of both dentofacial aesthetics and functional occlusion

[†] Equal contribution.
[*] Corresponding authors.

through teeth misalignment correction [13]. Its success relies heavily on high-quality 3D digital dental models for accurate diagnosis, treatment planning, and progress evaluation. Intra-oral scanning (IOS) has become the gold standard for high-precision 3D reconstruction [7], and advances in automatic teeth alignment have significantly improved clinical efficiency [22, 12, 3, 11, 20]. Nevertheless, orthodontic treatment typically spans a long period and requires frequent in-person monitoring. The high cost of equipment and time-consuming scanning process further restrict the accessibility of IOS, particularly in remote scenarios where a more flexible modeling solution is required.

In contrast, intra-oral photography offers a more accessible and cost-effective alternative, supported by the widespread availability of high-resolution mobile imaging devices. Reconstructing anatomically accurate 3D dental models directly from 2D intraoral images would fundamentally lower the threshold for digital orthodontics, enabling efficient remote diagnosis and fostering seamless patient-doctor communication. Consequently, 3D dental reconstruction from 2D images has emerged as a transformative research direction, promising to bridge the gap between high-end clinical precision and convenient, remote-ready dental care.

Achieving high-fidelity 3D reconstruction under constrained imaging conditions remains a formidable challenge in computer vision. Traditional geometric methods like SfM or SLAM [5, 18] rely on dense epipolar constraints that inherently fail under the sparse viewpoints of intra-oral photography. While recent implicit representations and 3D Gaussian Splatting have advanced view synthesis [14], they prioritize texture over geometric precision, yielding meshes unsuitable for clinical use. To mitigate this, statistical shape priors have been introduced for dental modeling. Chen et al. [1] employ a parametric teeth model for reconstruction via multi-parameter optimization. While capable of roughly reconstructing overall teeth shapes and arrangement, such methods lack strong anatomical priors for dental arches and are constrained by limited parameter spaces, often leading to reconstruction artifacts and considerable geometric errors. Moreover, the iterative multi-stage optimization often results in error accumulation and high inference latency. Recently, the success of diffusion models has inspired new paradigms in dental synthesis and reconstruction. TeethDreamer [23], for instance, leverages diffusion priors to augment sparse intra-oral views for subsequent surface reconstruction. Although this generative strategy alleviates data sparsity, the stochastic nature of diffusion may compromise cross-view consistency and geometric precision, particularly in unseen regions.

To address these challenges, we propose a framework called ARCH-Net to decompose the 3D teeth reconstruction task into three stages: (1) anatomically-constrained latent space construction, (2) counterfactual-enhanced 2D–3D alignment, and (3) prior-driven mesh refinement. First, we pretrain a Vector Quantized-Variational Autoencoder (VQ-VAE) [19] on 3D teeth point clouds to learn a robust latent space that captures the geometric distribution of dental anatomy. Within this space, a latent diffusion model bridges the gap between 2D contours and 3D geometry using cross-attention. We further introduce a counterfactual learning strategy that supervises the diffusion process with a con-

trastive loss between predicted latent codes and negative samples, enhancing the model's discriminative capability and ensuring precise geometric alignment. Finally, leveraging statistical teeth templates, we establish dense correspondences via sequential rigid and non-rigid registration, transforming coarse point clouds into high-resolution 3D meshes. Experimental results demonstrate that ARCH-Net achieves state-of-the-art performance, providing a high-fidelity solution for 3D dental reconstruction in clinical applications.

2 Method

An overview of the proposed ARCH-Net is illustrated in Fig. 1. In the first stage, we construct a latent embedding space by pretraining a VQ-VAE on 3D teeth point clouds (Sec. 2.1). In the second stage, we train a diffusion model in the learned latent space conditioned on 2D contours of intra-oral photographs to establish 2D-3D alignment so that the inferred latent code from 2D contours can be decoded into coarse point clouds (Sec. 2.2). In the third stage, with the statistical shape priors of teeth, we transform the predicted coarse point clouds into high-resolution 3D teeth meshes (Sec. 2.3).

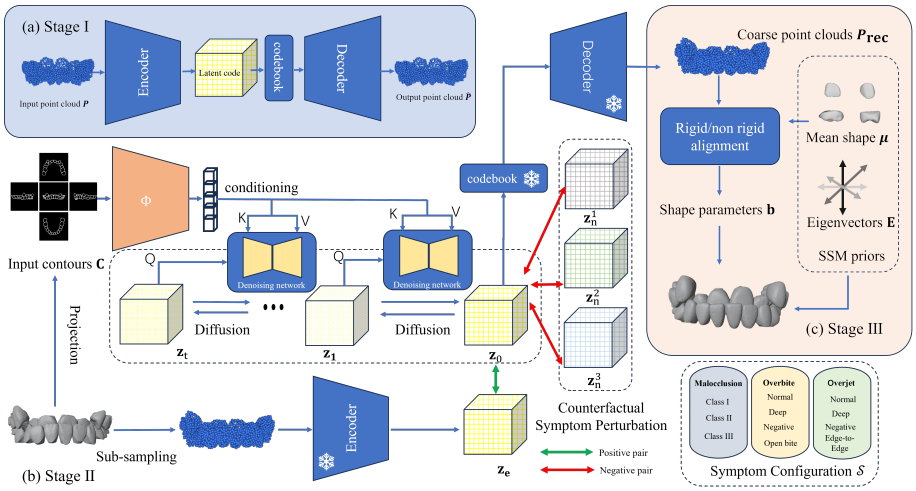

Fig. 1. Overview of ARCH-Net. (a) VQ-VAE pretraining on 3D teeth point clouds. (b) Counterfactual-enhanced 2D-3D alignment via conditional diffusion model. (c) Prior-driven mesh refinement.

2.1 Anatomically-constrained latent space construction

A straightforward approach to learning an end-to-end mapping from 2D images to a 3D point cloud is to encode the images into a latent representation and

decode it into 3D geometry, supervised by a geometric loss against the ground-truth point cloud. However, due to the inherent sparsity and ambiguity of 2D images, such a direct mapping strategy often fails to preserve high-level geometric semantics in the latent space, leading to limited generalization. As discussed in [4], an effective embedding should be both generative in 3D geometry and predictable from 2D image observations. Motivated by this principle, in this stage we first construct a geometrically structured generative latent space for 3D teeth point clouds.

Inspired by [10], we adopt VQ-VAE architecture to learn the latent representation. Specifically, let $\boldsymbol{P} \in \mathbb{R}^{K\times B\times 3}$ denote the input point clouds of K anatomically adjacent teeth, each represented by B points. The point clouds are encoded by an encoder $E(\cdot)$ to produce a latent embedding: $\mathbf{z} = E(\boldsymbol{P}) \in \mathbb{R}^{C\times H\times W\times D}$, where H, W, D denote compressed spatial dimensions and C is the embedding dimension. The continuous latent representation $\mathbf{z}$ is subsequently quantized using a learnable codebook in the VQ-VAE framework, yielding the discrete latent code $\mathbf{z}_q$. The decoder $D(\cdot)$ then reconstructs the teeth point clouds: $\hat{\boldsymbol{P}} = D(\mathbf{z}_q)$. To enforce geometric fidelity, we optimize the VQ-VAE using the Chamfer Distance (CD):

$$\mathcal{L}_{\text{rec}} = \sum_{\mathbf{x}\in\boldsymbol{P}} \min_{\hat{\mathbf{x}}\in\hat{\boldsymbol{P}}} \|\mathbf{x}-\hat{\mathbf{x}}\|_2^2 + \sum_{\hat{\mathbf{x}}\in\hat{\boldsymbol{P}}} \min_{\mathbf{x}\in\mathbf{P}} \|\hat{\mathbf{x}}-\mathbf{x}\|_2^2 . \tag{1}$$

2.2 Counterfactual-enhanced 2D-3D alignment

Conditional diffusion in latent space for 2D-3D alignment. Building upon the latent representation learned in Stage I, the second stage aims to establish alignment between 2D intra-oral photographs and 3D teeth models in the learned latent space. We formulate this alignment as a conditional diffusion probabilistic model that predicts the corresponding latent embedding by progressively denoising a Gaussian variable conditioned on 2D image features. Since our objective is to reconstruct the geometric shape and spatial arrangement of teeth rather than surface texture, we use tooth contours extracted from five intra-oral photographs as input: $\mathbf{C} = \left\{\mathbf{C}^{(s)}\right\}_{s=1}^{S}, \quad S = 5$, where $\mathbf{C}^{(s)} \in \mathbb{R}^{1\times h\times w}$ represents a single grayscale view. Let $\mathbf{f} = \Phi(\mathbf{C})$ be the feature representation extracted by an image encoder $\Phi(\cdot)$. Let $\mathbf{z}_e$ denote the latent embedding of the paired 3D teeth point clouds obtained from the frozen encoder $E(\cdot)$ of the pretrained VQ-VAE. Our objective is to model the conditional distribution: $p_\theta(\mathbf{z}_e \mid \mathbf{f})$, where θ denotes the parameters of the denoising network. During training, we define a Markov forward process that gradually adds Gaussian noise to the latent embedding $\mathbf{z}_e$: $q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right)$, where $\{\beta_t\}_{t=1}^{T}$ is a predefined variance schedule, $\mathbf{I}$ denotes the identity matrix, $\mathbf{z}_t$, $t = 0, 1, \ldots, T$ are latent variables and $\mathbf{z}_0 = \mathbf{z}_e$. We learn the reverse process conditioned on the image features: $p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{f}) = \mathcal{N}\left(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t, \mathbf{f}), \sigma_t^2\mathbf{I}\right)$. To effectively inject 2D information into the denoising process, instead of directly concatenating image features with timestep embeddings, we employ a cross-attention mechanism that aligns latent 3D tokens with 2D image features, enabling geometry-aware conditional generation.

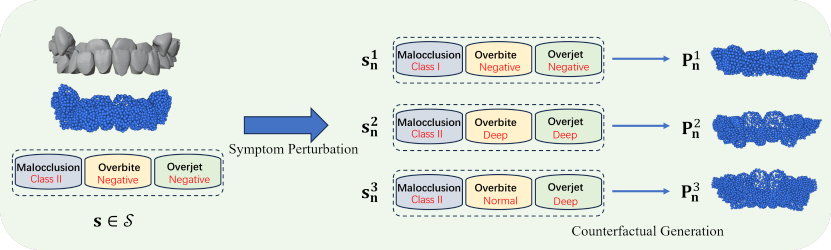

Fig. 2. Illustration of counterfactual generation via symptom perturbation.

Symptom-perturbation counterfactual learning. In orthodontics, teeth arrangement are characterized by clinically meaningful diagnostic attributes, such as overjet, overbite, and malocclusion. Because monocular 3D reconstruction from 2D images is inherently ill-posed, preserving these diagnostic characteristics is essential for anatomically and clinically plausible reconstruction. To encourage symptom-consistent latent representations, we introduce a counterfactual learning strategy. Specifically, let $\mathbf{s} \in \mathcal{S}$ denote the multi-label symptom vector associated with teeth point clouds $\boldsymbol{P}$. We synthesize counterfactual samples $\boldsymbol{P}_n$ whose symptom labels $\mathbf{s}_n$ differ from $\mathbf{s}$ in at least one diagnostic dimension (Fig. 2). Using the frozen VQ-VAE encoder, we extract latent embeddings $\mathbf{z}_e$ and $\mathbf{z}_n$ for $\mathbf{P}$ and $\mathbf{P}_n$, respectively. Given the 2D intra-oral contours $\mathbf{C}$, the diffusion model predicts the denoised latent representation $\mathbf{z}_\theta(\mathbf{z}_t, t, \mathbf{C})$. Beyond the standard diffusion reconstruction loss:

$$\mathcal{L}_{\text{diff}} = \|\mathbf{z}_0 - \mathbf{z}_\theta(\mathbf{z}_t, t, \mathbf{C})\|_2^2, \tag{2}$$

we introduce a counterfactual contrastive objective to enforce symptom-discriminative embeddings:

$$\mathcal{L}_{\text{counterfactual}} = -\log \frac{\exp\big(\mathbf{z}_\theta(\mathbf{z}_t, t, \mathbf{C})^\top \mathbf{z}_e/\tau\big)}{\sum_{z_n \in N} \exp(\mathbf{z}_\theta(\mathbf{z}_t, t, \mathbf{C})^\top \mathbf{z}_n/\tau)}, \tag{3}$$

where N denotes the set of counterfactual samples, τ denotes the temperature hyperparameter.

2.3 Prior-driven Mesh Refinement with Statistical Shape Models

In this stage, we incorporate a Statistical Shape Model (SSM) as a strong shape prior to transform the coarse teeth point clouds into dense meshes. Specifically, the pretrained VQ-VAE decoder takes the denoised latent code conditioned on 2D contours as input and generates coarse teeth point clouds $\boldsymbol{P}_{\text{rec}}$. For each tooth $v \in \mathcal{V}$, where $\mathcal{V}$ denotes the set of tooth labels, we first perform rigid alignment between the generated point cloud $\mathbf{P}_{\text{rec}}^v \in \mathbb{R}^{B\times 3}$ and the mean shape of the SSM $\boldsymbol{\mu}_v \in \mathbb{R}^{M\times 3}$ using the Coherent Point Drift (CPD) algorithm [16]. A low-rank Gaussian Process deformation model [1] is then applied to establish non-rigid correspondence, yielding an aligned dense point cloud $\mathbf{P}_{\text{aligned}}^v$. To enforce

anatomical plausibility, we compute the shape parameters by projecting it onto the SSM sub-space: $\mathbf{b} = \mathbf{E}_v^\top(\mathbf{P}^v_{\text{aligned}} - \boldsymbol{\mu}_v)$, where $\mathbf{E}_v \in \mathbb{R}^{3\times M\times J}$ denotes the top J principal eigenvectors obtained via Principal Component Analysis (PCA) [8] computed from a population of teeth shapes. Finally, the dense tooth geometry is reconstructed via back-projection: $\mathbf{P}^v_{\text{dense}} = \boldsymbol{\mu}_v + \mathbf{E}_v\mathbf{b}$.

3 Experiments

3.1 Experimental settings

Datasets. The experiments are conducted on a public 3D orthodontic dataset comprising 1,060 pre-treatment dental models [21], covering diverse clinical symptoms. Since the dataset does not include intra-oral photographs, we render five predefined views for each 3D model and extract teeth contours to simulate multi-view intra-oral contour images. Each contour image has a resolution of 256×256 pixels. The resulting dataset, consisting of paired 3D dental models and their corresponding multi-view contour images, is randomly divided into training, validation, and test subsets containing 800, 200, and 60 samples, respectively. To further evaluate the performance under real clinical conditions, we collect an additional 100 cases from a dental hospital. Each case includes real intra-oral photographs and the corresponding 3D intra-oral scan.

Implementation details Initially, we normalize the whole teeth model and sample 128 points from each tooth mesh using the Farthest Point Sampling (FPS) algorithm [15] to obtain coarse teeth point clouds. The VQ-VAE encoder consists of 32 tooth-wise PointNet++ modules [17] for local feature extraction, followed by a four-level 3D U-Net [2]. The decoder mirrors this structure and includes a four-level 3D U-Net with 32 tooth-wise linear heads for point coordinate regression. In Stage II, ResNet-18 [6] is adopted as the image encoder to extract contour features for conditional guidance. The denoising network is implemented with a 3D U-Net backbone augmented with four-head cross-attention layers to inject image features into the diffusion process. The number of diffusion time steps is set to 1000. For counterfactual sample construction, we synthesize three counterfactual samples for each training sample. Specifically, three samples with identical tooth composition but differing in only one diagnostic attribute are selected from the remaining dataset. Rigid registration is then performed between the original sample and the selected samples to simulate counterfactual cases. ARCH-Net is implemented in PyTorch and trained on a Linux platform with a single NVIDIA RTX 5880 Ada GPU. The batch size is set to 8 for both stages. Stage I and Stage II are trained for 500 and 1000 epochs, respectively, using the AdamW optimizer with learning rates of 1×10^{-4} and 1×10^{-3}.

We construct the statistical shape model (SSM) using Principal Component Analysis (PCA), retaining the first 20 principal components. For rigid alignment, we employ the Coherent Point Drift (CPD) algorithm with a maximum of 100

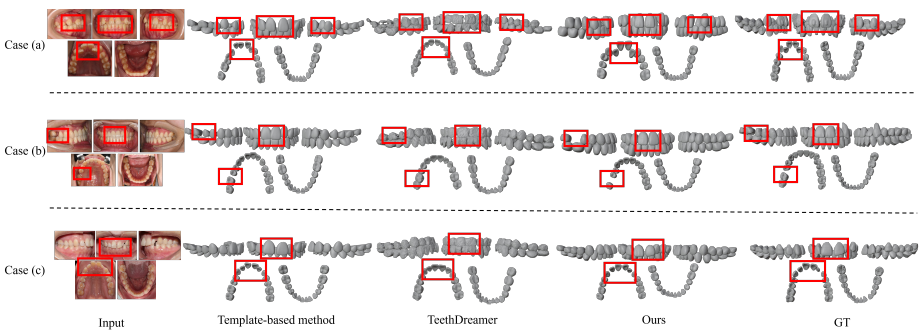

Fig. 3. Qualitative comparisons of reconstructed 3D teeth with other baselines using the real intra-oral photographs as input, demonstrating our results with complete shape and correct symptoms. (GT: ground truth)

iterations and a convergence tolerance of 10^{-4}. For Gaussian Process-based non-rigid deformation, we set the Gaussian kernel bandwidth to 10 and the kernel scale parameter to 2.

Baselines and metrics. We compare our approach with two representative baseline methods: an template-based framework proposed by [1] and TeethDreamer [23]. The selected template-based method represents the best-performing approach among existing template-based teeth reconstruction techniques, while TeethDreamer achieves state-of-the-art performance among learning-based methods. The template-based method accepts both synthesized teeth contours and real intra-oral photographs as input, whereas TeethDreamer operates directly on real photographs. In our experiments, we therefore use the collected real intra-oral photographs as input for both baselines. Since TeethDreamer employs the Segment Anything Model (SAM) [9] to segment teeth from intra-oral photographs as a preprocessing step, we also utilize SAM to extract teeth contours from intra-oral photographs for the template-based method and our proposed ARCH-Net to ensure a fair comparison.

As noted, our primary goal is to accurately reconstruct the geometric shape and spatial arrangement of the teeth, rather than texture or appearance. Accordingly, we evaluate the reconstructed meshes using root mean squared surface distance (RMSD), average symmetric surface distance (ASSD), Hausdorff distance (HD), Chamfer distance (CD), and Dice similarity coefficient (DSC).

3.2 Experimental results

Fig. 3 illustrates the 3D teeth reconstruction results of three typical cases using the real intra-oral photographs as input. The template-based method roughly recovers teeth morphology and arrangement but fails under severe malocclusion or irregular anatomy, sometimes producing artifacts such as inter-tooth collisions (Case (a) and Case (b) in Fig. 3). TeethDreamer captures fine-grained

Table 1. The quantitative comparison with other baselines and ablated solutions in teeth reconstruction.

Method	RMSD(mm)↓	ASSD(mm)↓	HD(mm)↓	CD(mm^2)↓	DSC↑
Template based method [1]	1.5466	1.1215	7.2298	6.3896	0.7641
TeethDreamer [23]	1.7346	1.3192	7.9312	6.6173	0.7415
Ours w/o cross attention	1.9326	1.4402	8.9633	7.5423	0.6593
Ours w/o counterfactual	1.8431	1.3711	8.5879	6.9463	0.7322
Ours	**1.3657**	**0.8103**	**6.3191**	**5.5132**	**0.7717**

occlusal grooves and surface patterns as evidenced by the upper and lower views, but yields suboptimal geometry and misses occlusal relationships since it reconstructs upper and lower arches separately. Our ARCH-Net addresses these limitations with a three-stage hierarchical design. Stage I leverages a pre-trained tooth point cloud embedding space to impose strong structural priors, preventing artifacts like tooth interpenetration. Stage II uses a diffusion-based 2D–3D alignment framework, enhanced with counterfactual learning, to improve teeth arrangement recovery, especially in severe misalignment cases (Case (a) in Fig. 3). Stage III performs prior-driven mesh refinement, enhancing mesh quality and morphological fidelity. In both Case (a) and Case (b), some teeth are missing compared to the ground truth due to limited visibility in the input intra-oral photographs. This demonstrates that our method naturally handles missing-tooth scenarios, which the other two baselines cannot. Overall, ARCH-Net achieves the best performance on real intra-oral photographs. Table. 1 presents quantitative results, consistent with the qualitative observations.

3.3 Ablation study

We conduct ablation studies to evaluate two key components in Stage II: cross-attention-based condition injection and counterfactual learning.

Quantitative results in Table. 1 (w/o cross-attention) and (w/o counterfactual) show that cross-attention significantly improves 2D–3D alignment. By allowing each 3D latent token to attend to semantically relevant regions in the 2D feature space, it establishes fine-grained correspondences during denoising. This mechanism effectively addresses the projection ambiguity inherent in intra-oral imaging; the attention-driven weighting scheme enables the framework to selectively suppress background noise and gingival interference, it ensures that critical anatomical landmarks, such as incisal edges and molar cusps, are precisely anchored to their 2D visual counterparts. Concurrently, the counterfactual learning strategy enforces symptom-consistent alignment in the latent space, improving geometric reconstruction quality while maintaining adherence to visual contour constraints. By simulating counterfactual scenarios during training through the targeted perturbation of specific dental features, the model is forced to decouple universal shape priors from patient-specific pathological traits. By learning to recognize these variations as distinct semantic entities, ARCH-Net maintains

high reconstruction fidelity even in complex cases where pathological features might otherwise be smoothed out by conventional generative priors.

4 Conclusion

In this paper, we present ARCH-Net, a hierarchical framework for reconstructing 3D dental models from five intra-oral photographs. By leveraging a pretrained tooth embedding space and counterfactual-enhanced 2D–3D alignment, our method effectively recovers overall tooth morphology and dental arrangement. A shape-prior-driven mesh refinement further improves reconstruction quality. Extensive experiments demonstrate that ARCH-Net outperforms existing methods. Future work will focus on investigating cross-view relationships among intra-oral images to further enhance structural consistency and reconstruction accuracy. We also plan to integrate recent advances in novel view synthesis to improve fine-grained texture and realistic tooth color, enhancing both geometric fidelity and visual realism.

Acknowledgments. This work was supported in part by Lingnan University Faculty Research Grant (106125), Lingnan Interdisciplinary & Strategic Research Grant (784009), SZU-LU Joint Research Programme (Grant No: SZU-LU009/2526) and submission control Macao Polytechnic University (fca.33bd.0c98.6)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, Y., Gao, S., Tu, P., Chen, X.: Automatic 3d teeth reconstruction from five intra-oral photos using parametric teeth model. IEEE Transactions on Visualization and Computer Graphics **30**(8), 4780–4791 (2023)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
3. Deng, Q., Yang, X., Huang, M., Jiang, L., Zhang, D.: Taposenet: Teeth alignment based on pose estimation via multi-scale graph convolutional network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 314–323. Springer (2024)
4. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: European conference on computer vision. pp. 484–499. Springer (2016)
5. Häming, K., Peters, G.: The structure-from-motion reconstruction pipeline–a survey with focus on short image sequences. Kybernetika **46**(5), 926–937 (2010)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hong-Seok, P., Chintal, S.: Development of high speed and high accuracy 3d dental intra oral scanner. Procedia Engineering **100**, 1174–1181 (2015)

8. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24**, 417–441 & 498–520 (1933)
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
10. Lei, C., Liang, Y., Wang, S., Dai, J., Liu, Y.J.: Teethgenerator: A two-stage framework for paired pre-and post-orthodontic 3d dental data generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25872–25881 (2025)
11. Lei, C., Xia, M., Wang, S., Liang, Y., Yi, R., Wen, Y.H., Liu, Y.J.: Automatic tooth arrangement with joint features of point and mesh representations via diffusion probabilistic models. Computer Aided Geometric Design **111**, 102293 (2024)
12. Li, X., Bi, L., Kim, J., Li, T., Li, P., Tian, Y., Sheng, B., Feng, D.: Malocclusion treatment planning via pointnet based spatial transformation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 105–114. Springer (2020)
13. Littlewood, S.J., Mitchell, L.: An introduction to orthodontics. Oxford university press (2019)
14. Miao, Y., Wu, T., Chen, T., Jiang, J., Tang, Z., Jiang, Z., Stefanidis, A., Yu, L., Su, J.: Dental3r: Geometry-aware pairing for intraoral 3d reconstruction from sparse-view photographs. 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) pp. 6086–6093 (2025)
15. Moenning, C., Dodgson, N.A.: Fast marching farthest point sampling. Tech. rep., University of Cambridge, Computer Laboratory (2003)
16. Myronenko, A., Song, X.: Point set registration: Coherent point drift. IEEE transactions on pattern analysis and machine intelligence **32**(12), 2262–2275 (2010)
17. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
18. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European conference on computer vision. pp. 501–518. Springer (2016)
19. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
20. Wang, C., He, Z., Chen, L., Yang, S., Sun, G., Duan, F., Wang, S., Zhou, Y.: Itmatch: Arch-guided semi-supervised tooth arrangement via iterative confidence evaluation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 268–277. Springer (2025)
21. Wang, S., Lei, C., Liang, Y., Sun, J., Xie, X., Wang, Y., Zuo, F., Bai, Y., Li, S., Liu, Y.J.: A 3d dental model dataset with pre/post-orthodontic treatment for automatic tooth alignment. Scientific data **11**(1), 1277 (2024)
22. Wei, G., Cui, Z., Liu, Y., Chen, N., Chen, R., Li, G., Wang, W.: Tanet: towards fully automatic tooth arrangement. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 481–497. Springer (2020)
23. Xu, C., Liu, Z., Liu, Y., Dou, Y., Wu, J., Wang, J., Wang, M., Shen, D., Cui, Z.: Teethdreamer: 3d teeth reconstruction from five intra-oral photographs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 712–721. Springer (2024)