Root Completion from Intraoral Scans of Tooth Crowns using Diffusion with Patch Perturbation

Yohan Jang
Korea University
Seoul, South Korea
dygks5412@korea.ac.kr

In-Seok Song
Korea University Anam Hospital
Seoul, South Korea
densis@korea.ac.kr

Seung Jun Baek*
Korea University
Seoul, South Korea
sjbaek@korea.ac.kr

Abstract

Intraoral scan (IoS) provides high-resolution data on the tooth crown, but does not contain information on the tooth root and thus has limitations in applications requiring 3D models of the whole tooth, e.g., virtual dental simulators. In this paper, we consider a diffusion-based model for root completion from IoS crowns. A key challenge is the lack of ground truth, i.e., the scan data of roots are typically unavailable. To train our model, we instead use the Cone-Beam CT (CBCT) data matched to IoS images, and use its crown as input and root as the pseudo-ground truth. Due to the difference in input data between training (CBCT crown) and inference (IoS crown), there is an issue of domain shift. To address the issue, we take a coarse-to-fine approach: we make a coarse prediction of roots using Coarse Estimator; introduce Perturbed Patch Generator (PPG) which generates patches from coarse points and perturbs them with noise for a robust prediction against the domain shift; and use Transformer denoiser for refined reconstruction. We also propose loss functions designed to facilitate the training of the denoiser with perturbed patches. Experiments show that our method outperforms prior techniques in various benchmark evaluations, demonstrating its robust performance in generating high-quality root data. Our code is available at https://github.com/yhJang94/RootCompletion.git.

1. Introduction

Intraoral scan (IoS) has become a standard practice in modern dentistry due to its ability to capture high-resolution details of the tooth crown in a non-invasive manner [15][1]. Unlike traditional impression-taking methods, IoS provides a digital representation of dental structures without causing discomfort or tissue damage. However, a major limitation of IoS is that it lacks information on the tooth root, which is essential for many applications [17][12][6]. For instance, VR/AR-based tooth extraction planning necessitates an accurate root structure to ensure safe and effective extraction procedures [20].

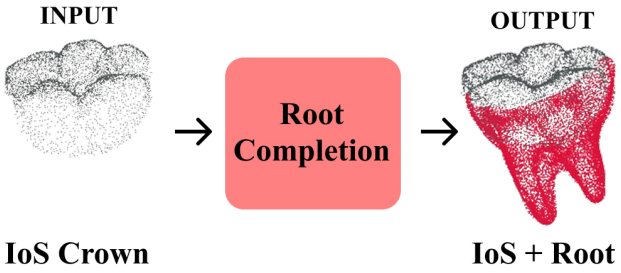


Figure 1. Generation of root structures from Intraoral scans (IoS) of crowns without ground-truth root data.

In this paper, we consider the problem of completing the root from IoS crowns based on diffusion models (see Fig. 1). The main difficulty is that the root scan data are hardly available, which implies the absence of ground truths to train diffusion models. To overcome the difficulty, we instead utilize Cone-Beam CT (CBCT) imaging which provides a 3D voxelized information on maxillofacial regions [5]. We prepare CBCT data of the subject matched to given IoS data, and segment the CBCT images to extract crown and root parts. CBCT crowns are used as input and CBCT roots are used as ground truth during training. IoS crowns are used during inference (completion). However, there exists a domain gap between input data for training and inference. For example, the crown from CT segmentation (training data) has imprecise boundaries, and mesh conversion from CT voxels for surface extraction causes point location errors relative to IoS crown meshes (inference data). In addition, we separate the crown and root of CT data to prepare the training data. The separation is based on the mesh registration to crown reference provided by IoS data, which is subject to registration error. Those factors of inaccuracies cause a domain shift, hampering the generalization capabil-

*Corresponding Author.

ity of the model.

In this study, we propose a novel framework for root completion using a Transformer-based diffusion model. We take a coarse-to-fine approach: a coarse prediction of roots is made from input crown, from which noisy patches are generated and refined with a denoiser. To make the model robust to domain shift, we propose the Perturbed Patch Generator (PPG). The PPG applies perturbations to the coarse estimate of roots, and the model is trained to reconstruct roots in the presence of perturbation. This makes our model resilient to estimation errors. Moreover, we introduce Patch-wise Denoising Loss (PDL) which concerns denoising patches of output point cloud selected to represent the overall geometric structure of the root. By focusing on denoising representative patches, PDL facilitates the optimization of our diffusion model, leading to improved quality of generation. Experiments show that our method outperforms the baseline models in both quantitative and qualitative aspects. Our contributions are summarized as follows: (i) we propose a novel framework for high-quality root completion from IoS crowns using Transformer-based diffusion; (ii) we develop Perturbed Patch Generator to make the model robust against the estimation errors caused by domain shift; (iii) we propose Patch-wise Denoising Loss which enhances structural consistency of generated roots by denoising patchified outputs.

2. Related Work

2.1. Point Cloud Completion.

Point completion aims to reconstruct the complete 3D shape from partial point clouds. PointNet [18] employs MLP independently on each point, achieving permutation invariance through global pooling. Building on this, PCN [30] introduced the first learning-based completion method with an encoder-decoder structure. Subsequent methods [21][26][23] have introduced hierarchical decoding to enhance reconstruction quality. For instance, TopNet [21] uses a tree-structured decoder to progressively expand point sets from latent features, while GRNet [26] converts point clouds into voxel grids, employing a coarse-to-fine strategy with 3D CNN and Gridding Residual Modules. Recent advancements have shifted towards transformer-based frameworks [24][28][31][27][29] to model long-range dependencies in point cloud completion. The skip-attention network [24] enhances feature transfer with skip connections between PointNet++ [19] encoders and decoders. PoinTr [28] transforms points into patches and uses a geometry-aware transformer for better structural capture, an idea extended by SeedFormer [31], which gradually generates fine output from patch seeds. AdaPoinTr [29] futher refines PoinTr with adaptive denoising queries to select meaningful coarse points. Point clouds from different methods can exhibit domain shifts, even with the same object. Our model focuses on recovering fine geometric details under such conditions.

2.2. Point Cloud Generators.

The generation of 3D points facilitates the expansion of point cloud processing into broader research domains. Initial generative models emerged from variational autoencoders (VAEs) [11] and generative adversarial networks (GANs) [8]. VRCNet [16] aligns input data with the ground-truth distribution using a shared encoder. Similarly, MM-Completion [25] maps input data to the target distribution with a pretrained VAE and refines the output through a GAN discriminator. Recently, diffusion-based models [14][4][10][3] have demonstrated performance in producing high-quality point clouds. For instance, PDR [14] employs a PointNet-based diffusion model for both coarse and fine stages. Other methods [4] introduce learnable rigid transformations to preserve structural integrity. These diffusion models are capable of generating high-quality data and exhibit inherent randomness. Consequently, some studies [25][3] propose distribution-based metrics for evaluation. Our approach uses a transformer-based diffusion model at the fine stage.

3. Preliminaries

3.1. Problem statement

Our goal is to complete the root part of the given IoS crown, for which we design a diffusion-based model. A key challenge is that the ground truth, the scanned root data, is typically unavailable. Alternatively, we prepare the matching CBCT data of the same subject as the IoS data. Our model is trained with the CBCT crown regarding the CBCT root as the pseudo-ground truth. However, there exists a domain gap in the input data between training (CBCT crown) and inference (IoS crown). The gap is not simply due to the difference in resolution; it arises mainly from the processing steps required to generate crowns and roots from CBCT data. Specifically, it results from (1) CT tooth segmentation errors, (2) mesh conversion errors for surface extraction, and (3) registration errors when separating CT teeth into crowns and roots. To quantify this gap, we measured the Fréchet point cloud distance between downsampled IoS crowns and registered CT crowns, obtaining 0.51. This is higher than the 2×10^{-4} observed between random downsamples of the same IoS crown, indicating a *domain shift*.

Also, our problem can be regarded as more ill-posed than typical point completion problems. A significant part (root) of the target for the completion (the whole tooth) is missing from the target. Thus, our completion task can be positioned in-between reconstruction and generation; but still our task requires that the generated root be smoothly connected to, and be anatomically consistent with, the input crown, which

poses a significant challenge.

3.2. Diffusion model

Recently, Denoising Diffusion Probabilistic Models (DDPM) [9] have been widely used for the generation of medical data. In this study, we use DiffPMAE [13] for root completion. DiffPMAE integrates a masked autoencoder (MAE) with diffusion processes, enabling the concurrent learning of geometric information and detailed local features of 3D objects. The MAE framework partitions the point cloud into two regions: the visible region and the masked region. Specifically, the crown region, denoted as x^c, is considered visible, while the root region, denoted as x^r, is treated as masked. During the forward diffusion process, Gaussian noise is incrementally applied solely to the root region. The noisy representation of the root at diffusion step t, denoted as x_t^r, is given by

$$x_t^r = \sqrt{\bar{\alpha}_t} x_0^r + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise, and $\bar{\alpha}_t$ is the cumulative product of noise scaling factors up to step t.

The decoder within the DiffPMAE framework is trained to predict the ground truth root data, enhancing the precision of generated root structures. The crown region remains unchanged, serving as geometric guidance for denoising. The decoder is trained to predict x_{t-1}^r based on noisy root x_t^r and the encoded crown patches $E(x_0^c)$ where t is the diffusion time step. The sampling of x_{t-1}^r from x_t^r is given by

$$x_{t-1}^r = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t^r + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_{\text{gen}}^r + \sigma_t x_{\text{gen}}^r. \tag{2}$$

where x_{gen}^r denotes the root predicted by the decoder based on x_t^r and $E(x_0^c)$, and $\alpha_t, \beta_t, \sigma_t$ are noise schedules at time step t.

4. Method

4.1. Root Completion Module

Overview. An overview of our method is shown in Fig. 2. We adopt a coarse-to-fine framework, where perturbations are applied to coarse points, and the output point cloud is refined based on the perturbed patches.

Coarse Estimator A set of patches is created from the input point cloud of a crown. The center points of patches are first sampled from input using Farthest Point Sampling (FPS), and the K-Nearest Neighbors (KNN) of a center point is set as the patch. We use this process to produce P patches ($P = 128$) where the number of points in a patch is K ($K = 32$). Next, a class embedding is computed based on the standard tooth number of the input crown, which identifies the tooth type. The patches and class embeddings are processed by a Transformer encoder, and we obtain the latent representation of P crown patches of dimension D ($D = 384$). A coarse prediction of the points of the tooth root is made based on the latent crown patches (Fig. 2) where an MLP is used to generate an initial estimate of N root points ($N = 384$).

Perturbed Patch Generator. The coarse points are processed by Perturbed Patch Generator (PPG) which we propose as a key module as follows. The coarse estimate of root points is obtained from CBCT crown during training, but will be from IoS crown during inference. Thus, the model will see a different distribution of coarse estimate at inference from that of training. We regard the estimate at inference as a distributionally shifted version of those seen during training. To make the model robust to distribution shifts, we deliberately *perturb* a subset of coarse estimates with random noise, and use the perturbed estimates to create patches which are used for refined reconstruction.

With the proposed module, our model is expected to learn to reconstruct the root in the presence of input perturbation. Note that the perturbation is applied only during training as shown in Fig. 3. The PPG module is designed as follows. Denote the points generated from the coarse estimator by $x_\theta \in \mathbb{R}^{N\times 3}$. We apply FPS to x_θ to extract a subset of M points ($M = 64$) from x_θ. The extracted points are perturbed by Gaussian noise with variance ρ^2 where ρ is a hyperparameter ($\rho = 0.01$). The M perturbed points are merged with the original N, and an MLP is applied to the merged points to compute the importance of each point. The top-N ranked points in importance, which are denoted by $\tilde{x}_\theta$, are selected. These points will be used as center points of patches. The rationale behind obtaining the perturbed points by applying FPS with Gaussian noise to the coarse estimate is as follows. Those operations make the perturbed points as a whole retain the shape of root from the coarse estimate and do not deviate excessively from the initial estimate. To guide the selection of points, we impose a loss $\mathcal{L}_{\text{coarse}}$ such that perturbed samples $\tilde{x}_\theta$ are close to a rough shape of the ground truth, or FPS samples of x_0^r denoted by $\text{FPS}(x_0^r)$, given by

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{CD}}(\tilde{x}_\theta, \text{FPS}(x_0^r)) \tag{3}$$

where $\mathcal{L}_{\text{CD}}$ is Chamfer Distance-L2 loss [7]:

$$\mathcal{L}_{\text{CD}}(X,Y) = \sum_{x\in X} \min_{y\in Y} \|x-y\|^2 + \sum_{y\in Y} \min_{x\in X} \|x-y\|^2 \tag{4}$$

Fine Generator. Patches of noisy root points x_t^r at diffusion time step t are generated using perturbed center points $\tilde{x}_\theta$. Let $\text{KNN}(x, y)$ denote the set of patches by applying KNN to points in x with respect to the patch center points in y. The patches are obtained by

$$\tilde{p}_\theta = \text{KNN}(x_t^r, \tilde{x}_\theta) \tag{5}$$

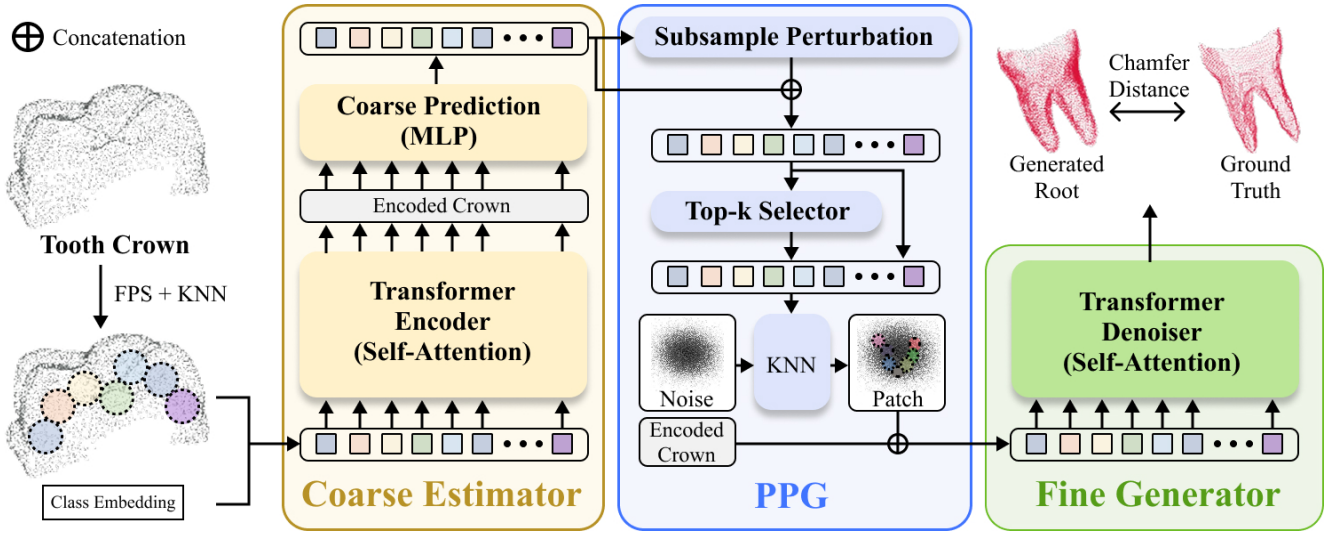


Figure 2. Overview of the proposed framework. The input partial points are divided into patches using Farthest Point Sampling (FPS) and K-Nearest Neighbor (KNN), then processed by a Transformer Encoder to extract latent features. A coarse estimator predicts the coarse points of roots based on latent features. We introduce Perturbed Patch Generator (PPG) to enhance model robustness and stability by adding a small perturbation into the coarse estimate during training. The generated patches are refined with a Transformer Denoiser to generate the fine details of the root.

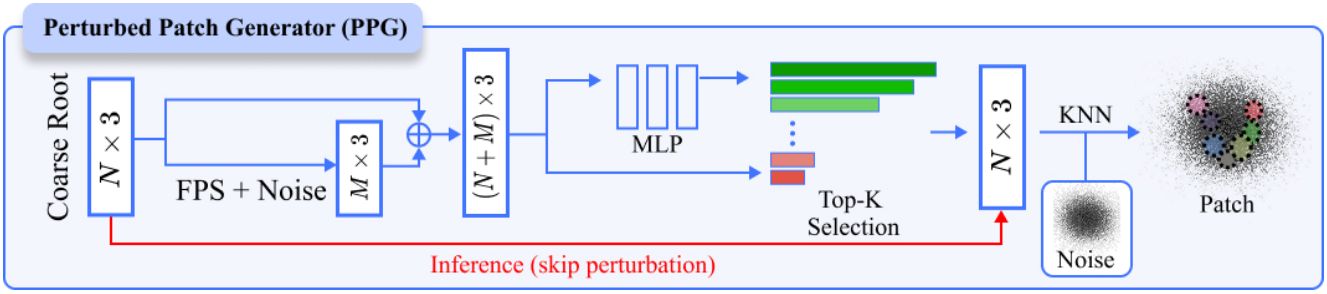


Figure 3. Illustration of the Perturbed Patch Generator (PPG), which mitigates domain shift between training and inference. During training, perturbations are added to the coarse root to enhance robustness. Inference proceeds without perturbations, ensuring stable predictions.

Based on perturbed coarse points, we obtain N patches ($N = 384$), where each patch consists of K neighboring points ($K = 32$). The root and crown patches are processed and input to a Transformer denoiser. The denoiser output denoted by x^r_{gen} is given by

$$x^r_{\text{gen}} = D(x^r_t, t, E(x^c_0), \tilde{p}_\theta) \tag{6}$$

For training, we use a denoising loss based on Chamfer Distance-L2 given by

$$\mathcal{L}_{\text{fine}} = \mathcal{L}_{\text{CD}}(x^r_{\text{gen}}, x^r_0) \tag{7}$$

For inference, x^r_{t-1} is estimated using Eq. (2) from x^r_t and x^r_{gen} in Eq. (6).

4.2. Optimization

Patch-wise Denoising Loss. We introduce a novel loss function for patch-based denoising. The overall denoising loss (7) focuses on the consistency of generated output in a point-wise manner. We propose that if we use patches which reflect the geometric features of the ground truth well and impose consistency on those patches, the optimization for denoising will be more focused and effective. The Patch-wise Denoising Loss (PDL) is proposed as follows. Given model output x^r_{gen} and ground truth x^r_0, we patchify each of them where the patch centers are given by FPS of x^r_0 so that the patches uniformly cover the ground truth to reflect its original shape well due to the property of FPS. The KNN is applied to the patch centers $\text{FPS}(x^r_0)$ to generate patches, and the Chamfer Distance-L2 is used to evaluate the consistency:

$$\mathcal{L}_{\text{PDL}} = \mathcal{L}_{\text{CD}}(\text{KNN}(x^r_{\text{gen}}, \text{FPS}(x^r_0)), \text{KNN}(x^r_0, \text{FPS}(x^r_0))) \tag{8}$$

Experiments show that PDL significantly improves the model performance.

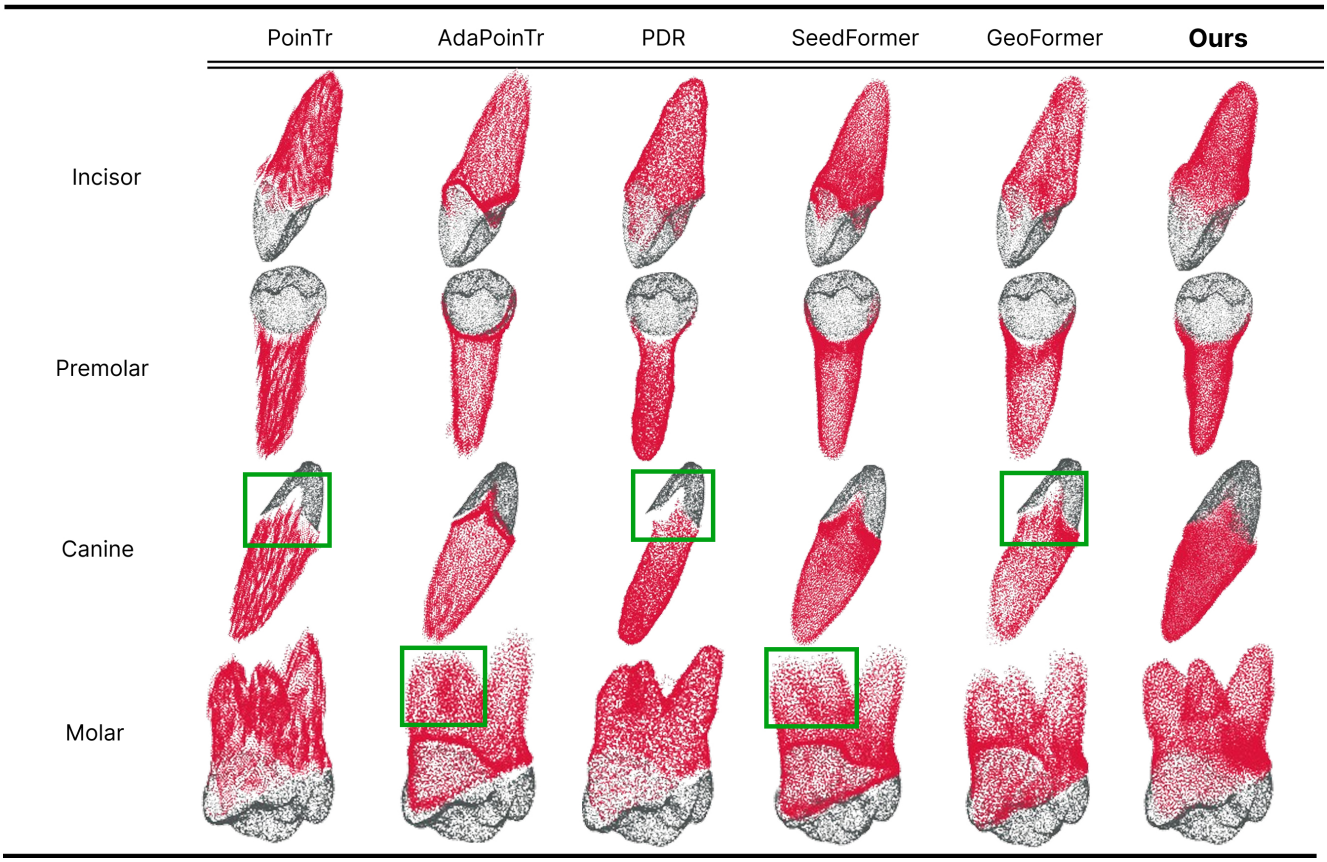


Figure 4. Qualitative comparison of generation of dental roots (red points) with the identical IoS crown (black points). The baseline models have issues with the generation (in green boxes) as follows: PoinTr and Geoformer generate roots along the crown's boundary but fail to maintain continuity. AdaPoinTr and Seedformer struggle to produce well-defined roots. The roots generated by PDR encroach on the crown or appear disconnected.

Repulsion Loss. We observed that, when excessive perturbation is added to the predicted coarse points in PPG, the overlapping regions between generated patches tend to grow large. Such clustering of patches is undesirable, because the patches should be uniformly distributed to cover and capture the overall shape. To address this issue, we use Repulsion Loss [22] to prevent excessive clustering and promote a more uniform distribution of points. The loss is given by

$$\mathcal{L}_{\text{repul}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_k(x_i)}(\text{ReLU}(\alpha - d_{ij}))^2 \qquad (9)$$

where N is the number of points, $\mathcal{N}_k(x_i)$ is k-nearest neighbors of x_i $(k=1)$, and α is a predefined repulsion threshold $(\alpha = 0.05)$. In conclusion, the overall loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{fine}} + \mathcal{L}_{\text{PDL}} + \lambda \cdot \mathcal{L}_{\text{repul}} \qquad (10)$$

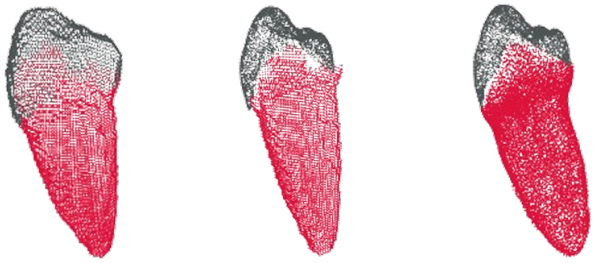
where λ is a hyperparameter $(\lambda = 0.1)$ for balancing losses.

5. Experiment

5.1. Experimental Setup

Dataset. This study obtained 3,282 tooth samples from 143 anonymized pairs of CBCT and IoS scans from (anonymized) institution. The samples were divided into 2,502 CBCT scans for training, 224 for validation, and 556 IoS scans for testing. Institutional Review Board approval was obtained (IRB No. (anonymized)). The CBCT images have a resolution of $768 \times 768 \times 576$ with a voxel size of $0.3 \times 0.3 \times 0.3$ mm^3. IoS scans were obtained using Medit i500 scanner in an in-vivo resolution of 50 μm. Each tooth was independently annotated by two experts and verified by an oral and maxillofacial surgeon. We aligned the paired CBCT and IoS data using the Iterative Closest Point (ICP) algorithm [2] and split the CBCT teeth into crown and root regions. The crown and root were downsampled to 4,096 and 12,288 points, respectively, via FPS.

Baselines. We compare our model with the following baseline methods for point completion. PoinTr [28] uses a



CT Crown + CT Root | **IoS Crown + CT Root** | **IoS Crown + Inference**

Figure 5. A visualization of crown-root pairs across different domains. The discrepancy in modality and registration errors often result in a mismatch between the CT root and the IoS, especially near boundaries (center). In contrast, the predicted root forms a more natural and anatomically plausible structure (right).

geometry-aware transformer with KNN-based relations and DGCNN feature aggregation. AdaPoinTr [29] estimates missing points through a dynamic query bank derived from input points and encoder outputs. SeedFormer [31] introduces Patch Seeds to encode positional and feature information for local geometries. PDR [14] is a diffusion model designed with a PointNet-based dual-path architecture. GeoFormer [27] incorporates 2D multi-view images and partial point clouds for prediction. DiffPMAE [13] is excluded from the baseline, because it requires a coarse set of ground truth points within the region of reconstruction; in our task, the whole root region is missing.

Evaluation Metrics. We consider *two* types of metrics for quantitative evaluation due to the mismatch between modalities as follows. The inference output from our model is a root completed naturally from the given IoS (right, Fig. 5). However, the CT root which is used as a (pseudo-) ground truth, does not have smooth boundaries with the IoS crown due to the difference in modalities (center, Fig. 5). Thus, a direct evaluation of the reconstruction accuracy of the generated roots against CT roots as a whole is inappropriate due to the modality gap near the crown-root boundaries. We instead consider two types of metrics: (i) *distribution-based* metrics, comparing the distributions between generated and CT roots, are utilized to assess the overall quality of generated outputs; (ii) *reconstruction accuracy* of generated roots against CT roots is measured after removing the root points near the crown-root boundaries to exclude the effect of inherent modality gap in the data. Below, we describe the metrics in detail.

For distribution-based metrics, we use Maximum Mean Discrepancy (MMD), Coverage (COV), and 1-Nearest Neighbor Accuracy (1-NNA) based on Chamfer Distance (CD) and Earth Mover's Distance (EMD). MMD quantifies distributional difference, COV measures the diversity of reference matches, and 1-NNA assesses leave-one-out accuracy. Additionally, Jenson-Shannon Divergence (JSD) is utilized to evaluate differences between probability distribution, leveraging the Kullback-Leibler divergence. For metrics of reconstruction accuracy, standard measures such as CD, EMD, and the 95% Hausdorff Distance (HD95) are adopted to evaluate point-wise and structural similarity. CD calculates the average closest-point distance between two point sets, EMD determines the minimal cost of transforming one distribution into another, and HD95 assesses large deviations while mitigating the influence of outliers.

Method	MMD (↓)		JSD (↓)	1-NNA (↓)	COV (↑)	
	CD	EMD		CD	CD	EMD
PoinTr	1.58	17.22	4.20	99.91	30.04	15.65
SeedFormer	0.22	12.67	<u>3.90</u>	39.93	70.86	13.49
PDR	3.55	9.94	44.94	100.00	3.60	4.14
AdaPoinTr	0.22	17.42	8.22	<u>36.24</u>	76.08	12.41
GeoFormer	<u>0.21</u>	<u>9.24</u>	7.59	37.50	<u>79.32</u>	<u>18.88</u>
Ours	**0.14**	**6.09**	**2.21**	**20.86**	**84.17**	**27.34**

Table 1. Quantitative comparison of IOS generation results in distribution-based metrics. MMD and JSD are scaled by 100; 1-NNA and COV are reported in %. The best and second-best performances are highlighted in **bold** and <u>underline</u>, respectively.

5.2. Experimental Results

Quantitative Results. Table 1 presents a quantitative comparison between the proposed and baseline models in *distribution-based* metrics. Among the baseline methods, Transformer-based models [29][31][27] without PoinTr [28] demonstrate commendable performance on metrics such as MMD, 1-NNA, JSD. In particular, AdaPoinTr [29] achieves its performance through an adaptive mechanism. SeedFormer [31] improves completion by progressively refining local geometries from coarse points. GeoFormer [27] appears to enhance completion by incorporating multi-view 2D features extracted from canonicalized point clouds. In contrast, CD-based COV metrics generally exhibit high coverage, whereas EMD is more sensitive to misalignments, resulting in lower scores. This discrepancy is likely due to slight registration errors between the IoS and CT, which may cause directional deviations in the generated roots. Meanwhile, PoinTr [28] achieved favorable scores in MMD and JSD, indicating effective maintenance of the overall distribution. However, its low 1-NNA score reveals a deficiency in accurately predicting root morphologies corresponding to the crown input, often resulting in the generation of a single or averaged root shape. In another case, PDR [14], employing a coarse-to-fine strategy with a PointNet-based [18] diffusion model, showed poor performance across all metrics. This may be due to the independent processing of each point in PointNet [18] which hinders its ability to capture fine-grained structures of IoS data compared to Transformer-based models. Overall, the

Method	CD-L1 (↓)	CD-L2 (↓)	EMD (↓)	HD95 (↓)
PoinTr	0.171	0.0163	0.328	0.507
SeedFormer	0.099	0.0025	0.170	0.525
PDR	0.137	0.0493	0.440	0.589
AdaPoinTr	0.127	0.0025	0.224	0.454
GeoFormer	0.155	0.0024	0.186	0.594
Ours	**0.022**	**0.0016**	**0.070**	**0.061**

Table 2. Quantitative comparison of the reconstruction accuracy of the root completion.

issue of domain shift may have negatively affected the performance of the baseline models. Our results show that the proposed model outperformed the baselines in all evaluation metrics, demonstrating its robustness to domain gaps.

Table 2 shows the *reconstruction accuracy* of the generated roots against the ground truth (CT) roots. The comparisons of generated and CT roots were made after removing the root points within a normalized distance of 0.03 from the crown; this is to exclude the effect of modality gap near the crown-root boundaries for a fair comparison. SeedFormer [31], AdaPoinTr [29], and GeoFormer [27] generally exhibited favorable performance. Conversely, PoinTr [28] and PDR [14] scored below the average in CD-L1. The performance disparity became more pronounced in CD-L2, which is more sensitive to outliers. A similar decline was observed in the EMD score, indicating potential issues with distortion or point density, as reflected in the qualitative results. In the HD95 metric, models demonstrated performance comparable to other baselines due to the exclusion of the top 5% outliers. Our model consistently outperformed others by generating coarse predictions deterministically and refining them stably through a transformer-based diffusion model, showing its ability to generate root structures closely resembling those of CT, despite the domain shift.

Qualitative Results. Figure 4 presents qualitative comparisons for dental root generation from identical IoS crown inputs. Our model generates anatomically plausible root structures across various tooth types, including incisors, canines, premolars, and molars. It produces smooth surfaces, natural curvature, and clear bifurcations, especially in complex multi-rooted molars. In contrast, PoinTr [28] generates sparse and unevenly distributed points, often resulting in fragmented shapes with noticeable outliers. Despite using a geometry-aware transformer, it struggles with IoS inputs. PDR [14] exhibits discontinuities at the crown-root boundary and occasionally intrudes into the crown. These issues likely stem from its point-wise and stochastic training paradigm, which hinders robustness under domain shift. AdaPoinTr [29], SeedFormer [31], and GeoFormer [27] produce relatively cleaner structures with fewer artifacts. However, they often fail to capture fine local geometry such as subtle curvature and detailed bifurcations. This limitation likely comes from the inherent constraints of transformer-based architectures. In contrast, our approach facilitates learning global features with coarse estimation followed by the refinement of local details through patch-based diffusion processing. Our approach enables producing continuous tooth structures with clear and well-defined roots aligned with the crown.

Method	MMD(↓)		JSD(↓)	1-NNA(↓)	COV(↑)	
	CD	EMD		CD	CD	EMD
Base	0.24	9.57	3.94	36.02	77.53	19.26
w/o perturb	0.22	8.61	2.60	33.36	79.86	22.12
Ours	**0.14**	**6.09**	**2.21**	**20.86**	**84.17**	**27.34**

Table 3. Ablation study of network components. 'Base' indicates training without a PPG, while 'w/o perturb' refers to training without perturbation, but with patch generation.

Method	MMD(↓)		JSD(↓)	1-NNA(↓)	COV(↑)	
	CD	EMD		CD	CD	EMD
Base	0.17	8.11	5.10	29.95	80.40	18.90
w/o PDL	0.15	7.12	4.95	28.24	80.76	19.78
w/o Repul	0.15	6.32	2.69	20.93	82.32	23.02
Ours	**0.14**	**6.09**	**2.21**	**20.86**	**84.17**	**27.34**

Table 4. Ablation study of loss terms. The 'base' model is trained without PDL and repulsion loss. PDL preserves local geometric details, while repulsion loss reduces point clustering, collectively enhancing performance.

Ablation Study. Table 3 presents an ablation study evaluating the effectiveness of the Perturbed Patch Generator (PPG) within our coarse-to-fine framework. The base model directly refines root structures without perturbation or patch generation from coarse points. In patch-based setting, coarse points are used to perform KNN-based grouping and also serve as positional encodings. Without this alignment, the base model performs refinement and shows low performance. Without perturbation, the model achieves moderate performance, comparable to other baseline methods, but shows limited generalization under domain shifts. Our full model integrates both perturbation and patch-based alignment, enabling precise root completion and achieving the best performance across all evaluation metrics. These findings demonstrate that PPG contributes significantly to robust completion under domain discrepancies.

Table 4 shows that both Patch-wise Denoising Loss (PDL) and repulsion loss play important roles in achieving accurate and stable completion. The base model, which excludes both PDL and repulsion loss, struggles to preserve fine geometric details and tends to produce locally clustered points. It yields the lowest performance across all evaluation metrics. Without PDL, the model fails to capture fine-grained geometry within each patch, causing overall performance degradation. In the absence of repulsion loss, the model generates locally clustered points due to perturbation applied to coarse points. Since the perturbed coarse points

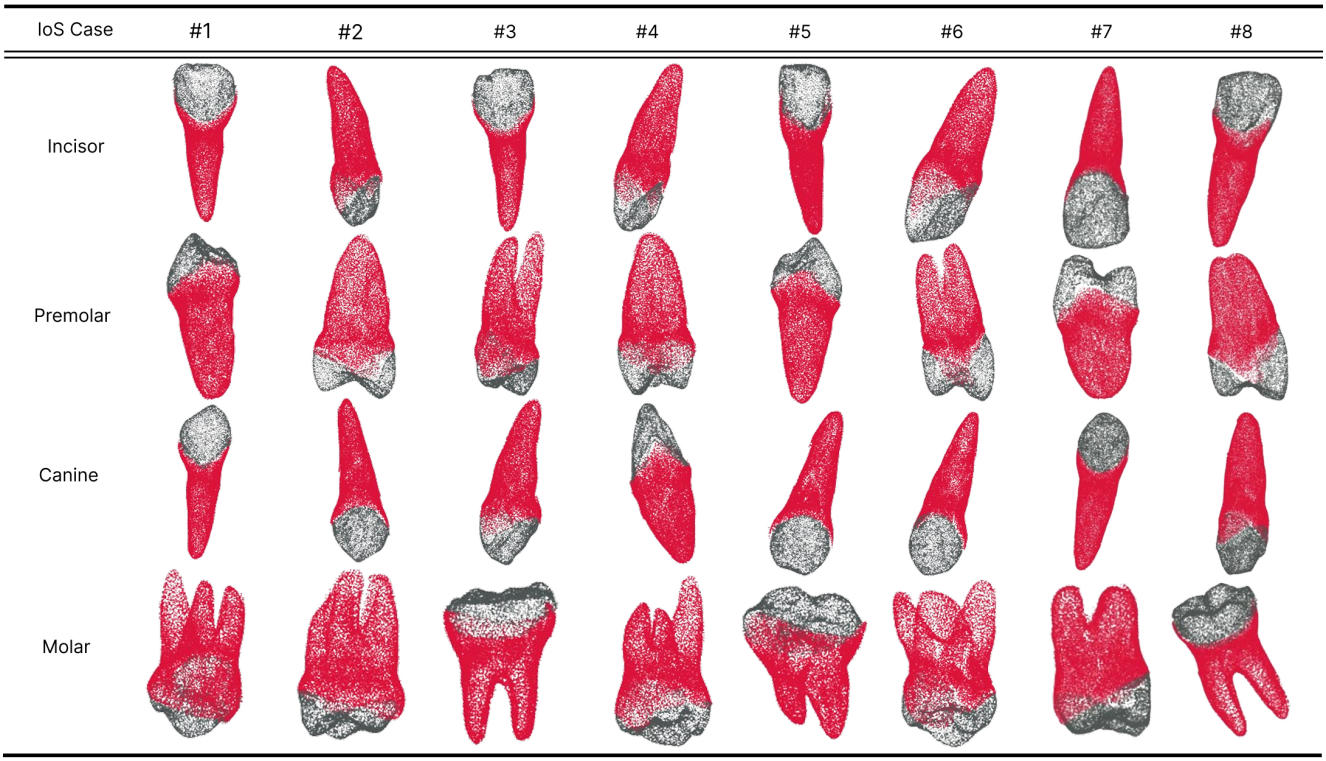


Figure 6. Qualitative results of root generation across various IoS. The model predicts roots according to crown morphology. For incisors and canines, it predicts a single root. Premolars typically have one root, except for the maxillary first premolar, where the model predicts two. For molars, it predicts three roots for maxillary and two for mandibular molars. Fused roots are also observed in columns 2 and 7·

are used for patch generation, this can lead to overlapping regions between neighboring patches. These overlaps cause non-uniform point distributions, leading to noticeable performance drops in COV. In contrast, MMD, JSD, and 1-NNA remain relatively unaffected. The full model achieves the best results across all evaluation metrics.

Anatomical Analysis. Figure 6 illustrates the qualitative outcomes of root generation from various IoS. Our model effectively generates root structures that align with the crowns of incisors and canines. For example, our model generates a single root for the premolar, but produces two distinct roots for maxillary teeth numbered 14 and 24. This behavior indicates that the model recognizes the anatomical difference in crown morphology specific to tooth position. A similar pattern is observed in molars, where the model generates three roots for maxillary and two for mandibular. Occasionally, it generates fused roots in the maxillary molars. This suggests the model captures anatomical variations such as root fusion, observed in the training data. Notably, tooth fusion was confirmed in the ground truth of the same case, offering valuable insights for clinical procedures like root canal treatment. This observation underscores the model's ability to internalize anatomical priors from partial 3D inputs. While recent studies focus on recovering 3D structures from 2D images, such as panoramic X-rays, few address root reconstruction from partial crowns like IoS. By incorporating insights from both modalities, future research could enhance the accuracy and clinical relevance of 3D reconstructions.

6. Conclusion

In this study, we propose a Transformer-based diffusion model for the generation of dental roots from IoS crowns. We introduce Perturbed Patch Generator (PPG) which enhances the robustness of our model to the domain shift problem caused by the modality gap. In addition, we propose Patch-wise Denoising Loss (PDL) that performs a patch-level optimization to facilitate the denoising process and generate precise root details. Experimental results show that our method generates naturally connected roots with crowns compared to the baseline approaches. Our future research includes integrating additional information, such as 2D panoramic X-ray images, for 3D reconstruction and generation of tooth models viable for clinical applications.

Acknowledgements. This work was supported by the ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201819), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2022-NR070834).

References

[1] Pekka Ahlholm, Kirsi Sipilä, Pekka Vallittu, Minna Jakonen, and Ulla Kotiranta. Digital versus conventional impressions in fixed prosthodontics: a review. *Journal of Prosthodontics*, 27(1):35–41, 2018. 1

[2] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992. 5

[3] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2272, 2023. 2

[4] Jisheng Chu, Wenrui Li, Xingtao Wang, Kanglin Ning, Yidan Lu, and Xiaopeng Fan. Digging into intrinsic contextual information for high-fidelity 3d point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2573–2581, 2025. 2

[5] W De Vos, Jan Casselman, and GRJ19464146 Swennen. Cone-beam computerized tomography (cbct) imaging of the oral and maxillofacial region: a systematic review of the literature. *International journal of oral and maxillofacial surgery*, 38(6):609–625, 2009. 1

[6] Mihaela Dutã, Corneliu I Amariei, Crenguta M Bogdan, Dorin M Popovici, Nicolae Ionescu, and Cristina I Nuca. An overview of virtual and augmented reality in dental education. *Oral Health Dent Manag*, 10(1):42–9, 2011. 1

[7] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[10] Yoni Kasten, Ohad Rahamim, and Gal Chechik. Point cloud completion with pretrained text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 12171–12191, 2023. 2

[11] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, page 307–392, 2019. 2

[12] Yaning Li, Hongqiang Ye, Fan Ye, Yunsong Liu, Longwei Lv, Ping Zhang, Xiao Zhang, and Yongsheng Zhou. The current situation and future prospects of simulators in dental education. *Journal of Medical Internet Research*, 23(4): e23635, 2021. 1

[13] Yanlong Li, Chamara Madarasingha, and Kanchana Thilakarathna. Diffpmae: Diffusion masked autoencoders for point cloud reconstruction. In *ECCV*, 2024. 3, 6

[14] Zhaoyang Lyu, Zhifeng Kong, Xudong Xu, Liang Pan, and Dahua Lin. A conditional point diffusion-refinement paradigm for 3d point cloud completion. *arXiv preprint arXiv:2112.03530*, 2021. 2, 6, 7

[15] Francesco Mangano, Andrea Gandolfi, Giuseppe Luongo, and Silvia Logozzo. Intraoral scanners in dentistry: a review of the current literature. *BMC oral health*, 17:1–11, 2017. 1

[16] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu. Variational relational point completion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8524–8533, 2021. 2

[17] Philipp Pohlenz, Alexander Gröbe, Andreas Petersik, Norman Von Sternberg, Bernhard Pflesser, Andreas Pommert, Karl-Heinz Höhne, Ulf Tiede, Ingo Springer, and Max Heiland. Virtual dental surgery as a new educational tool in dental school. *Journal of cranio-maxillofacial surgery*, 38(8): 560–564, 2010. 1

[18] 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. 2, 6

[19] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 2017. 2

[20] Marcus Rieder, Bernhard Remschmidt, Christina Gsaxner, Jan Gaessler, Michael Payer, Wolfgang Zemann, and Juergen Wallner. Augmented reality-guided extraction of fully impacted lower third molars based on maxillofacial cbct scans. *Bioengineering*, 11(6):625, 2024. 1

[21] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 383–392, 2019. 2

[22] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7774–7783, 2018. 5

[23] Xiaogang Wang, Marcelo H. Ang Jr. , and Gim Hee Lee. Cascaded refinement network for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. Point cloud completion by skip-attention network with hierarchical folding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[25] Rundi Wu, Xuelin Chen, Yixin Zhuang, and Baoquan Chen. Multimodal shape completion via conditional generative adversarial networks. In *Computer Vision–ECCV 2020: 16th*

European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pages 281–296. Springer, 2020. 2

[26] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun. Grnet: Gridding residual network for dense point cloud completion. In *European conference on computer vision*, pages 365–381. Springer, 2020. 2

[27] Jinpeng Yu, Binbin Huang, Yuxuan Zhang, Huaxia Li, Xu Tang, and Shenghua Gao. Geoformer: Learning point cloud completion with tri-plane integrated transformer. In *ACM Multimedia 2024*, 2024. 2, 6, 7

[28] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, 2021. 2, 5, 6, 7

[29] Xumin Yu, Yongming Rao, Ziyi Wang, Jiwen Lu, and Jie Zhou. AdaPoinTr: Diverse Point Cloud Completion With Adaptive Geometry-Aware Transformers . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pages 14114–14130, 2023. 2, 6, 7

[30] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 international conference on 3D vision (3DV)*, pages 728–737. IEEE, 2018. 2

[31] Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. Seedformer: Patch seeds based point cloud completion with upsample transformer. In *Computer Vision – ECCV 2022*, pages 416–432, 2022. 2, 6, 7