Tooth Sequence Transformer: Multi-Tooth Context Attention for Functional and Anatomical Tooth Reconstruction

Dissertation
zur Erlangung des Doktorgrades
der Humanwissenschaften
(Dr. sc. hum.)

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Alexander Broll
aus
Regensburg

im Jahr
2026

Dekan:	Prof. Dr. Dirk Hellwig
Betreuer:	Prof. Dr. Sebastian Hahnel
Mentor:	Prof. Dr. Martin Rosentritt
Mentor:	Prof. Dr. Markus Goldhacker
Tag der mündlichen Prüfung:	01. 07. 2026

Abstract

Objectives: Fully automated reconstruction of anatomically detailed and functionally compatible tooth crowns from partial digital impressions remains challenging. This work aims to develop and evaluate a deep-learning-based pipeline that reconstructs functional, patient-specific crown morphology from multi-tooth context and leverages case-specific parametric optimization to achieve contact-aware placement within the dental arch.

Methods: We propose Tooth Sequence Transformer (TST), a tooth-sequence-driven transformer that operates on segmented 3D tooth point clouds. The data is represented in a normalized adjacent teeth frame and conditioned on tooth identity via anatomical priors. The network predicts a dense target tooth using adjacent and antagonist context, followed by Tooth Pose Predictor (TPP) for initial rigid alignment and a Tooth Pose Optimizer (TPO) stage that refines translation and scale under occlusal and proximal contact objectives. The pipeline is evaluated against common point cloud completion baselines using morphological and contact-specific metrics as well as qualitative reconstructions.

Results: Across anterior and posterior tooth ranges, TST achieves statistically significant improvements in morphological fidelity and proximal fit compared to both baselines, including lower chamfer distance and adjacent tooth contact losses. For posterior teeth, TST reduces antagonist penetration relative to the baselines. While the baselines can generate plausible tooth morphology for anterior teeth, only TST yields plausible and high-fidelity reconstructions for posterior teeth, as confirmed by qualitative assessment. The TPO stage significantly improves adjacent contact alignment, reducing the adjacency loss by approximately a factor of 5–6 compared to pre-optimization poses. Ablation studies confirm the impact of key model design choices.

Conclusions: The proposed TST pipeline provides functional crown proposals that combine learned morphology with controllable, contact-driven placement refinement. The method outperforms common point cloud completion baselines in morphological accuracy and qualitative fidelity, particularly for morphologically complex posterior molar teeth.

Zusammenfassung

Ziele: Die vollautomatische Konstruktion anatomisch detaillierter und funktionell gestalteter Kronen aus partiellen digitalen Abformungen bleibt eine anspruchsvolle Aufgabe in der Zahnmedizin. Ziel dieser Arbeit war die Entwicklung und Evaluation eines Deep-Learning-basierten Verfahrens, das unter Berücksichtigung von Antagonisten, Nachbarzähnen und den restlichen Zähnen des Zahnbogens eine gleichermaßen funktionelle und patientenspezifische Kronenmorphologie rekonstruiert und über eine fallspezifische parametrische Optimierung eine kontaktpunktbezogene Positionierung innerhalb des Zahnbogens ermöglicht.

Methoden: Wir präsentieren einen zahnsequenzgetriebenen Transformer (Tooth Sequence Transformer (TST)), der segmentierte 3D-Zahnpunktwolken in einem normalisierten, an den Nachbarzähnen ausgerichteten Koordinatensystem verarbeitet und über anatomische Vorgaben auf die Zahnidentität konditioniert wird. Das Netzwerk rekonstruiert die Form der Krone im Nachbar- und Antagonistenkontext. Anschließend wird eine initiale translatorische Starrkörpertransformation berechnet (Tooth Pose Predictor (TPP)). Darauf aufbauend werden Translation und Skalierung anhand okklusaler und approximaler Verlustfunktionen optimiert (Tooth Pose Optimizer (TPO)), um Zahnkontakte gezielt zu berücksichtigen. Die Methode wird gegenüber etablierten Verfahren zur Punktwolkenrekonstruktion mittels morphologischer und kontaktspezifischer Metriken sowie anhand qualitativer Ergebnisse evaluiert.

Ergebnisse: Über Front- und Seitenzähne hinweg erreicht TST gegenüber beiden Baselines statistisch signifikante Verbesserungen sowohl in den morphologischen Metriken als auch in der approximalen Passung. Dies zeigt sich unter anderem in einer geringeren Punkt zu Punkt Distanz sowie in reduzierten Kontaktverlusten zu benachbarten Zähnen. Für Seitenzähne verringert TST zudem die Durchdringung mit Antagonisten signifikant im Vergleich zu den Baselines. Während die Baselines bei Frontzähnen eine plausible Morphologie erzeugen, liefert nur TST auch für Seitenzähne konsistente, plausible und detailgetreue Rekonstruktionen. Die TPO-Stufe verbessert insbesondere die Approximalkontakte deutlich und reduziert die zugehörige Verlustgröße im Mittel um etwa den Faktor 5–6 gegenüber dem initialen Modelloutput. Ablationsstudien stützen die zugrundeliegenden Modell- und Pipeline-Entscheidungen.

Schlussfolgerungen: TST generiert funktionelle Kronendesigns, indem es eine datengetrieben gelernte Morphologie mit einer gezielt steuerbaren, kontaktbasierten Optimierung der Translation und Skalierung kombiniert. Im Vergleich zu generischen Verfahren erreicht die Methode, insbesondere bei morphologisch komplexen Seitenzähnen, eine höhere morphologische Genauigkeit und eine bessere qualitative Detailtreue.

Contents

List of Figures

List of Tables

List of Symbols

Mathematical Notation

$\mathbf{0}$	All-zeros matrix.
$\boldsymbol{A}^{\text{abs}}$	Matrix with the element-wise absolute values of $\boldsymbol{A}$.
$\boldsymbol{A}^{-1}$	Inverse of matrix $\boldsymbol{A}$.
$\boldsymbol{A}^{\text{T}}$	Transposed matrix $\boldsymbol{A}$.
$\boldsymbol{A}$	Matrix.
$\mathbb{1}$	All-ones matrix.
$\mathbb{I}$	Identity matrix.
$\boldsymbol{\nabla}(f)$	Gradient of f.
$\boldsymbol{x}^{\text{T}}$	Row vector.
$\boldsymbol{x}$	Column vector.
$\mathcal{A}$	Set.
$\text{Attn}_{\text{cross}}$	Cross-attention operator.
Attn	Attention operator.
CD	Chamfer Distance operator.
IoU	Intersection over Union operator.
LN	Layer normalization operator.
concat	Concatenation operator.
eig	Eigendecomposition operator.
$a \in \mathcal{A}$	Element a of set $\mathcal{A}$.
x^*	Optimum of x.
x	Scalar.

Mathematical Symbols

C'	PosePredictor embedding width.
C_{H}	Extended TST embedding width.
C	TST Embedding width.
M	Cluster centers of occluded points (contact points).
Q	Number of query tokens.
Δ	Offset function.
Π	Query generator.
α	Rotation angle.
β	Generic weighting factor.
$\mathbf{\Lambda}$	Eigenvalue matrix.
$\mathbf{\Sigma}$	Covariance matrix.
$\boldsymbol{\kappa}$	Curvature value.
n_{t}	Number of teeth.
κ	Curvature value.
λ	Eigenvalue.
λ	Loss weight.
$\mathbb{N}$	Set of natural numbers.
$\mathbb{R}$	Set of real numbers.
$\mathbf{H}$	Stacked feature matrix.
$\mathbf{N}$	Point cloud normals matrix.
$\mathbf{P}$	Point cloud matrix.
$\mathbf{R}$	Rotation matrix.
$\mathbf{V}$	Eigenvector matrix.
$\mathbf{X}$	Centered point cloud matrix.
$\mathbf{Z}$	Query matrix.
$\mathbf{e}$	Unit vector.
$\mathbf{v}$	Eigenvector.
$\mathcal{D}$	Set of Euclidean distances.
$\mathcal{F}$	FDI index set of teeth.
$\mathcal{L}$	Loss function.
$\mathcal{N}$	Set of point cloud normals.
$\mathcal{P}$	Point cloud - set of 3D points.
$\mathcal{S}$	Set of point clouds.

$\mathcal{T}$	Set of per-tooth translations.
$\mathcal{U}$	Set of per-tooth transforms.
$\mathcal{Z}$	Query set.
ν	Orientation indicator function.
ω	Prediction head.
ϕ	Transformer Encoder/Decoder function.
ψ	Embedding function.
ρ	Projection from features to coarse points.
φ	Linear fusion map.
ϑ	Threshold parameter.
$\boldsymbol{\delta}$	Offset.
$\boldsymbol{d}$	Direction vector.
$\boldsymbol{e}$	Embedding vector.
$\boldsymbol{h}$	Feature vector.
$\boldsymbol{m}$	Cluster center of occluded points (contact point).
$\boldsymbol{n}$	point cloud normal vector.
$\boldsymbol{p}$	Point in 3D space.
$\boldsymbol{q}$	Unit quaternion.
$\boldsymbol{r}$	Orientation reference vector.
$\boldsymbol{t}$	Translation vector.
$\boldsymbol{u}$	Transformation vector.
$\boldsymbol{z}$	Query vector.
ξ	Contact-point encoder.
c	Arbitrary coefficient or factor.
d	Euclidean distance.
f	FDI index of tooth.
n	Count of elements.
p	Penetration scalar.
q	FDI quadrant index.
s	Isotropic scale factor.
t	FDI tooth index within a quadrant.
v	Vicinity scalar.

Acronyms

AI	artificial intelligence.
CAD	computer aided design.
CD	chamfer distance.
CPG	coarse point generator.
CPM	contact point module.
DGCNN	dynamic graph convolutional neural network.
DL	deep learning.
FDI	Fédération Dentaire Internationale.
FPS	furthest point sampling.
GAN	generative adversarial network.
GT	ground truth.
kNN	k-nearest neighbors.
MISE	multi-resolution iso-surface extraction.
MLP	multilayer perceptron.
NKSR	neural kernel surface reconstruction.
PCA	principal component analysis.
PCD	point cloud.
PEM	progressive expansion module.
SGD	stochastic gradient descent.
TPO	Tooth Pose Optimizer.
TPP	Tooth Pose Predictor.
TST	Tooth Sequence Transformer.

1 Introduction

Recent advances in artificial intelligence (AI) have accelerated the digital transformation of dentistry [1–4]. In parallel, generative models specifically tailored to dental applications have been introduced [5, 6], including approaches that explicitly target restorative dentistry [7–9] with the aim of enabling accurate and efficient prosthesis design.

Conventional manual restoration workflows remain labor-intensive and are subject to substantial operator-dependent variability. Computer aided design (CAD) and manufacturing systems mitigate these limitations by reducing manual workload and improving the consistency of reconstructions. Chairside solutions further provide an integrated pipeline that combines intraoral scanning, fixed dental prosthesis design, and fabrication into a single streamlined workflow for smaller restorations. This provides dental practices with an efficient and technically accessible end-to-end solution.

Although conventional CAD software provides an initial restoration proposal, substantial manual refinement is still often required to satisfy prosthetic constraints and case-specific functional demands, again relying on the clinician's expertise. Preserving functional dynamic contacts during mandibular excursions as well as static contacts in terminal occlusion remain central to this refinement process. Integration of motion-tracking software supports patient-specific articulation by capturing individual mandibular movements [10, 11], thereby enabling functional occlusal assessment and subsequent manual adaptation of the geometry within the CAD environment. Despite these capabilities, manual adjustment remains prone to errors due to the complexity and patient-specific variability of the masticatory system, compounded by the need to respect material and fabrication-dependent minimal material thickness requirements [12]. Inaccuracies may result in occlusal discomfort, temporomandibular disorders, or even restoration failure [13, 14].

Clinical adjustments prolong treatment time and increase overall cost, while corrective material removal may weaken the restoration, induce surface damage, and elevate the risk of fracture [14]. These limitations highlight the potential of fully automated deep

learning (DL)-based reconstruction pipelines as an alternative to traditional tooth libraries and rule-based occlusal schemes. Such approaches assume that patient-specific functional and morphological characteristics of the restoration are implicitly encoded in the observed dentition and surrounding structures, and can therefore be learned and inferred by data-driven models. Previous research, as summarized in the following section, has demonstrated that generative DL architectures can synthesize anatomically plausible and functionally compatible tooth morphology directly from 3D scans, thereby reducing reliance on manual adjustments and expert knowledge. Based on the acquired 3D data of the patient's dentition, the input can be represented as depth images, volumetric grids, surface meshes, or point clouds (PCDs) and can thus be classified as 2D or 3D approaches.

1.1 State of the art

This section reviews prior work on data-driven dental crown reconstruction and transformer-based 3D PCD completion. To provide a comprehensive overview of generative approaches, it covers both domain-specific dental reconstruction frameworks and generic transformer architectures for 3D PCD completion. This highlights methodological trends and key architectural design principles. The literature research strategy is detailed in the appendix (Section A.1).

1.1.1 Dental reconstruction using generative deep learning

Data-driven tooth reconstruction methods can be categorized into approaches that operate on 2D depth or intensity images and methods that process 3D geometry directly [15].

2D-approaches. Early generative pipelines adopt a 2D formulation in which the prepared and opposing teeth are projected into depth images, enabling the reuse of standard image-to-image architectures. Hwang et al. [16] generate depth maps of the missing crowns by projecting the prepared quadrant and the opposing teeth into 2D representations and feeding them into a pix2pix [17] generative adversarial network (GAN). The data is additionally augmented with loss terms that encode the gap distances between occluding surfaces. Yuan et al. [18] extend this concept through conditional generative models, that operate on occlusal-depth images synthesized from 3D scans, introducing

perceptual losses and occlusal-groove filters to better preserve fissure morphology and antagonist relationships. Tian et al. [19–21] present a series of adversarial frameworks that remain entirely within the 2D domain. Initially, a Wasserstein-based inlay restoration network with local and global discriminators for crown-defect completion is proposed. This is extended by a dual-discriminator architecture emphasizing biologically plausible occlusal morphology, and the two-stage DCPR-GAN. DCPR-GAN first reconstructs a coarse occlusal surface and subsequently refines grooves and occlusal fingerprints using additional constraints. Gu et al. [22] concentrate on local pit-and-fissure morphology by training conditional GANs to produce high-detail occlusal depth images that are then used to deform an existing crown mesh. Building on StyleGAN models for image synthesis, Broll et al. [23] project single molar tooth meshes into depth maps, train a generative image model on complete occlusal views, and apply a downstream Bayesian optimization strategy for partial inlay reconstruction. More recently, Roh et al. [24] integrate a segmentation network with a GAN-based inpainting module to generate crown depth images at varying positions in the dental arch.

Collectively, these 2D approaches leverage mature convolutional architectures and benefit from computational efficiency. However, the 2D projection step inevitably discards geometric information. As a result, these methods are limited in their capacity to capture undercuts and the full three-dimensional relationships among larger contiguous tooth segments, antagonists, and surrounding structures.

3D-approaches. To mitigate the information loss inherent in depth-map projections, more recent work has shifted toward representations that operate directly on 3D geometry. At the single-tooth level, Ye et al. propose VF-Net [25], which learns probabilistic latent representations of dental PCDs for mesh generation and unsupervised shape completion. Chanintonsongkhla et al. [26] employ PointFlow to model the distribution of anterior tooth PCDs, enabling both unconditional sampling and latent-code-based reconstruction of corrupted crowns.

Early generative approaches for tooth synthesis based on remaining teeth employ volumetric or PCD-based GANs. Chau et al. [27] evaluate a 3D adversarial framework trained on digital casts for single-molar prostheses. Ding et al. [28] adopt a 3D-DCGAN to learn crown morphology and compare its geometric and mechanical performance with biogeneric designs generated by conventional CAD software. Lessard et al. [29] adapt a multi-resolution adversarial completion network to synthesize crown PCDs from incom-

plete contextual geometry comprising the prepared tooth, adjacent teeth, and antagonists.

The geometry aware PCD completion transformer PoinTr proposed by Yu et al. [30, 31] has inspired several dental reconstruction frameworks. Zhu et al. [32] propose a PoinTr-based [30] framework, which integrates a geometry-aware transformer encoder with a dedicated surface-reconstruction module to complete missing teeth on single jaw models and generate watertight meshes. Ma et al. [33] follow a similar approach using the AdaPoinTr [31] model to reconstruct a single incisor based on the five surrounding teeth of the upper jaw. Hosseinimanesh et al. [34] introduce Dental Mesh Completion (DMC), a transformer-driven point-to-mesh generator trained on local adjacent and antagonist PCD context. The method is trained on various target teeth. The PCD prediction transformer module is again based on PoinTr [30]. They subsequently extend this line of work to a dedicated point-to-mesh completion network incorporating contrastive chamfer losses and margin-line supervision for clinically oriented crown design [35]. DCrownFormer and its extended mesh-refinement variant combine morphology-aware transformer blocks with differentiable surface-reconstruction layers to generate high-fidelity crown meshes from scans of preparations, adjacent teeth and antagonists [36, 37].

Beyond PCD-transformer architectures, Ping et al. [38] utilize self-attention implicit-function networks that jointly complete crowns and gingival structures from partial inputs, demonstrating that implicit 3D representations can recover fine dental morphology. Saleh et al. [39] compare transformer- and diffusion-based generative models for reconstructing missing teeth in full arches. Wu et al. [40] couple an image-restoration network with an enhanced Pixel2Mesh module [41] to reconstruct 3D meshes of defective teeth from single RGB inputs. Liu et al. [42] introduce TUCNet for high-resolution PCD completion of sparse dental scans. Du Chen et al. [43] present ToothDIT, which deforms an implicit template via a learned latent code to restore anterior incisor morphology. Feng et al. [44] jointly estimate pose and shape of maxillary anterior crowns via coupled transformer networks operating on crown PCDs. They highlight the importance of disentangling morphology and positioning in 3D generative reconstruction and respecting the sequential organization of teeth along the dental arch.

Across both 2D and 3D methods, existing work demonstrates that generative models can reconstruct anatomically plausible crowns with error levels that are competitive with, or in some cases superior to, conventional CAD-based workflows. A summary of the reviewed methodologies is provided in Table 1.1.

Table 1.1: Summary of the reviewed DL-based dental reconstruction methodologies.

Reference	Network	Data representation
	2D	
Hwang et al. [16]	GAN	depth map
Yuan et al. [18]	GAN	depth map
Tian et al. [19]	GAN	depth map
Tian et al. [20]	GAN	depth map
Tian et al. [21]	GAN	depth map
Gu et al. [22]	GAN	depth map
Broll et al. [23]	GAN	depth map
Roh et al. [24]	GAN	depth map
	3D	
Ye et al. [25]	Variational Autoencoder	point cloud
Chanintonsongkhla et al. [26]	PointFlow	point cloud
Chau et al. [27]	GAN	voxel
Ding et al. [28]	GAN	voxel
Lessard et al. [29]	GAN	point cloud
Zhu et al. [32]	Transformer	point cloud
Ma et al. [33]	Transformer	point cloud
Hosseinimanesh et al. [34]	Transformer	point cloud
Hosseinimanesh et al. [35]	Transformer	point cloud
Yang et al. [36]	Transformer	point cloud
Yang et al. [37]	Transformer	mesh
Ping et al. [38]	Implicit function network	voxel
Saleh et al. [39]	Transformer / Diffusion	deep implicit function / mesh
Wu et al. [40]	Pixel2Mesh	mesh
Liu et al. [42]	Graph convolutional network	point cloud
Du Chen et al. [43]	DeformNet	signed distance field
Feng et al. [44]	Transformer	point cloud

1.1.2 Transformers for 3D point cloud completion

Recent dental reconstruction methods increasingly leverage transformer architectures for 3D PCD completion, capitalizing on their ability to model long-range dependencies and complex geometric relationships [32, 34, 36, 37, 44]. These models are often based on architectures originally developed for generic object-level completion tasks. Therefore, the domain specific articles often lack architectural detail and design rationale. To provide a comprehensive overview of transformer-based 3D PCD completion, key architectural components and design principles from recent literature are summarized. The research is categorized into proxy- and seed-based methods, path-based refinement and feedback

networks, and attention-only or diffusion-based networks. A preliminary overview of point cloud completion is provided to contextualize the transformer-based approaches.

Point cloud completion. Point cloud completion aims to infer a complete 3D shape from a partial observation. Early progress in learning directly from unordered 3D coordinates was enabled by PointNet and its extensions [45, 46], which established permutation-invariant feature learning and inspired a broad range of downstream 3D understanding tasks. Building on this basis, PCN [47] became a widely used completion baseline by introducing an encoder–decoder design that first predicts a coarse set of completed points and then refines it to a denser output. Its refinement stage follows a FoldingNet-style [48] strategy that learns to deform a regular 2D point set into a 3D surface patch, providing an efficient mechanism to upsample the coarse prediction. Subsequent work, as outlined in the following paragraphs, has largely focused on improving resolution, robustness, and handling of noisy partial inputs, forming the methodological basis for later transformer-based completion models.

Proxy- and seed-based geometry-aware transformers. PoinTr [30] reformulates PCD completion as a set-to-set translation problem by grouping the partial input into local patches, mapping each patch to a point proxy with learned positional features, and processing these proxies with a geometry-aware transformer encoder-decoder. The encoder models global relationships among partial proxies, while the decoder predicts a sparse set of complete-shape proxies that are subsequently mapped to dense PCDs in a coarse-to-fine manner using chamfer distance (CD) losses for supervision. AdaPoinTr [31] builds on this framework with an adaptive query generation mechanism that conditions decoder queries on the specific partial shape and introduces a denoising-based completion task. Robustness and diversity are improved while reducing computational cost. ProxyFormer [49] refines the proxy formulation by explicitly distinguishing existing and missing-part proxies, introducing a missing-part-sensitive transformer that operates entirely in proxy space. This is supported by a proxy-alignment loss that makes missing proxies explicitly responsive to the geometry of the occluded regions. SeedFormer [50] replaces complete-shape proxies with patch seeds, learning a compact seed representation and an upsample transformer that progressively upsamples these seeds into dense completions while interpolating seed features to finer points. More recently, PoinTr-PM [51] augments the PoinTr backbone with a multi-stage point-moving decoder that iteratively displaces noise

points around predicted centers, reducing completion noise and improving fine-detail reconstruction compared to FoldingNet-style [48] upsampling.

Path-based refinement and feedback completion networks. Complementary transformer completion frameworks focus on multi-stage refinement rather than exclusively on proxy design. PMP-Net++ [52] interprets completion as moving points along learned paths from partial observations toward the complete shape. It uses transformer-enhanced feature encoders and multiple displacement stages to recover global structure first and progressively refine local details. FBNet [53] combines a hierarchical graph-based encoder-decoder that produces a coarse completion with a feedback refinement module built from feedback-aware completion blocks. At each refinement stage, the network re-injects the partial input and previous predictions to iteratively correct errors and densify the output. CRA-PCN [54] explicitly targets cross-resolution feature aggregation through intra- and inter-level cross-resolution transformer blocks, enabling coarse global context and fine-scale detail features to interact throughout a coarse-to-fine upsampling pipeline.

Attention-only and diffusion-based networks. Several works explore transformers that operate directly on point tokens, seeds, or voxel patches without hand-crafted neighborhood constructions. ShapeFormer [55] proposes a transformer-based shape completion framework that encodes partial shapes into sparse latent representations and decodes them into complete shapes, effectively leveraging transformers for sparse-to-dense reconstruction via learned shape tokens. PointAttN [56] demonstrates that attention alone can serve as a strong backbone for PCD completion by processing per-point tokens with stacked attention blocks that jointly capture local geometric details and global shape context, avoiding explicit k-nearest neighbors (kNN)-based grouping while achieving competitive or superior reconstruction accuracy. DiT-3D [57] adapts diffusion transformers to 3D shape generation by voxelizing PCDs into 3D grids and applying windowed transformer blocks with 3D positional encodings within a diffusion process. This yields high-fidelity and diverse 3D shapes and demonstrates that large plain transformers can effectively scale to 3D geometry.

Taken together, the transformer-based completion models establish a rich set of design principles, including proxy or seed tokenization of partial shapes, geometry-aware attention mechanisms, coarse-to-fine upsampling, multi-stage point-moving or feedback refinement, and diffusion-based generative modeling.

1.1.3 Limitations of existing approaches

Across both the dental reconstruction literature and generic transformer-based PCD completion, many approaches demonstrate that anatomically plausible tooth shapes can be synthesized from partial inputs. However, translating these advances into a clinically usable crown-design pipeline requires more than geometric completion. In particular, restorative design requires tooth-type awareness, patient-specific functional context, and a placement mechanism that can be adjusted to satisfy occlusal and proximal-contact requirements while respecting geometric constraints of the restorative workflow.

A recurring limitation is the restricted effective context used by many learning-based approaches. Several methods are primarily conditioned on the prepared tooth and immediate neighbors or formulate the task as isolated single-tooth completion. As a result, antagonist information and broader multi-tooth context are often omitted or only weakly encoded, even though they are essential to infer patient-specific occlusal contacts and to maintain consistent proximal target relationships along the dental arch.

Limitations also arise from the underlying data representations and training objectives. Early 2D depth-image pipelines rely on projections that discard 3D information needed to resolve undercuts and complex cusp–fossa morphology. Even when operating directly on 3D geometry, many methods combine general PCD completion architectures with generic shape similarity objectives such as CD or Earth Mover's distance, which emphasize overall geometric agreement but do not directly encode anatomical or functional requirements.

Finally, existing transformer-based dental models only partially exploit the structured nature of the dentition. While some works embed the target position in an arbitrary sequence or rely on local positional encodings [37, 44], they do not explicitly combine global quadrant information with local Fédération Dentaire Internationale (FDI) tooth-index features to model anatomical identity and arch position jointly. This restricts how morphology and functional context can be shared across tooth types and jaw configurations and, to our knowledge, there is no dedicated sequence-based method that targets posterior tooth reconstruction at the level of full functional crown design. In parallel, generic transformer completion architectures are typically developed and evaluated on object-centric, category-level benchmarks and thus do not natively incorporate dental anatomy, occlusion, or workflow-specific constraints, which can limit their direct applicability to restorative settings.

1.2 Summary

These limitations jointly motivate a reconstruction framework that operates directly on segmented 3D tooth PCDs and represents the dentition as a structured sequence with explicit FDI encoding and quadrant-aware positional information. The pipeline couples a transformer-based generative architecture with downstream pose prediction and optimization stages that enforce occlusal and proximal contact constraints as well as minimum thickness requirements for clinically viable crown design.

We propose Tooth Sequence Transformer (TST), a novel 3D PCD-based, tooth-sequence-driven reconstruction pipeline for single-tooth surface design in both anterior and posterior regions. The method operates on segmented 3D tooth PCDs within a normalized adjacent teeth coordinate frame and explicitly encodes tooth identities according to the FDI numbering system, capturing quadrant and tooth-index information as well as adjacent and antagonist context.

The transformer-based generative reconstruction network predicts the target tooth from multi-tooth context in a coarse-to-fine manner. Each tooth is represented as a sequence element via a dynamic graph convolutional neural network (DGCNN)-based [58] tooth encoder. The encoder extracts morphological and anatomical quadrant-tooth-index FDI features as well as functional occlusal features through a contact point module (CPM) from the input PCDs. Thus, kNN-based grouping is not required for local point proxy construction as commonly implemented in PCD-transformers for general or dental-specific reconstruction networks [30, 32, 34, 36, 37]. After feature enrichment via attention-based encoding, a coarse point generator (CPG) infers a coarse target that is subsequently upsampled to a dense reconstruction using a two-stage transformer decoder with a progressive expansion module (PEM).

Using the same tooth encoder architecture, the Tooth Pose Predictor (TPP) leverages positional and anatomical information from the tooth sequence, together with attention-based feature enrichment and an multilayer perceptron (MLP), to predict an initial rigid transformation for the reconstructed dense target tooth. The final Tooth Pose Optimizer (TPO) module refines this initial pose by optimizing a composite objective that integrates occlusal contact constraints with antagonists, proximal contact constraints with adjacent teeth, and minimal-thickness requirements relative to a prepared tooth surface. To support accurate evaluation of the optimization objective, the TPO employs neural kernel surface reconstruction (NKSR) [59] to generate a high-resolution mesh from the

reconstructed target PCD, enabling precise assessment of contact and thickness metrics on the final surface.

Together, these components constitute the TST pipeline, which is designed to produce crowns that are both morphologically coherent and functionally compatible with patient-specific occlusal and proximal relationships.

2 Materials and methods

This section details the proposed TST pipeline for functional 3D tooth reconstruction as illustrated in Fig. 2.1. It is organized as follows:

- Section 2.1 summarizes the data representations and common operations used throughout this work.
- Section 2.2 describes the overall data processing and normalization pipeline.
- Section 2.3 details the transformer-based generative reconstruction network TST.
- Section 2.4 presents the TPP module.
- Section 2.5 describes the TPO procedure.
- Section 2.6 defines the evaluation metrics.
- Section 2.7 outlines the experimental setup.

2.1 Data and function representation

Common operations and data representations used throughout this work are summarized in the following section. This includes the used indexing scheme for teeth via FDI identifiers, the representation of 3D tooth PCDs as sets or stacked matrices, and the representation of per-tooth transformations.

2.1.1 Indexing and FDI identifier

The teeth are ordered according to their position in the dental arch that is encoded via the FDI numbering system. A tooth identifier is given by a quadrant-tooth representation $f = 10\,q + t$, where $q \in \{1, 2, 3, 4\}$ and $t \in \{1, \ldots, 7\}$. Teeth with $t > 7$ are not considered in the final tooth sequence due to their limited presence in the dataset. The set of all

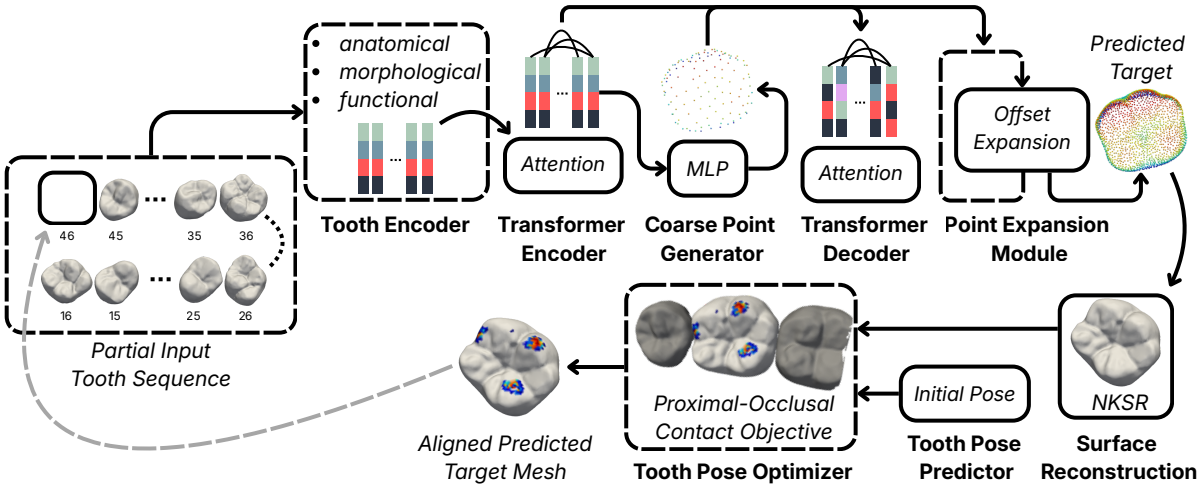


Fig. 2.1: Overview of the proposed TST reconstruction pipeline. Segmented 3D tooth PCDs are represented in a normalized frame and encoded as a tooth sequence with anatomical and functional conditioning. The transformer-based generative reconstruction network predicts a dense target tooth from multi-tooth context in a coarse to fine manner. The mesh surface is reconstructed using NKSR. A subsequent Tooth Pose Predictor module estimates an initial rigid transformation, which is refined by the Tooth Pose Optimizer stage that optimizes translation and scale under an occlusal and proximal contact objective.

FDI indices for a patient is denoted as $\mathcal{F}$. Indexing operations in this work will refer to a tooth via its FDI identifier f.

2.1.2 Object representation

A PCD is defined as a permutation invariant set of points

$$\mathcal{P} = \left\{ \boldsymbol{p}_i \mid \boldsymbol{p}_i \in \mathbb{R}^3 \ , \ i \in \{1, \dots, |\mathcal{P}|\} \right\} = \{\boldsymbol{p}_i\}_{i=1}^{|\mathcal{P}|} \subset \mathbb{R}^3$$

with the cardinality $|\mathcal{P}|$ as the number of points. The point set can be equivalently represented as a matrix

$$\mathbf{P} \in \mathbb{R}^{|\mathcal{P}| \times 3}, \qquad \mathbf{P} = \begin{bmatrix} \boldsymbol{p}_1^\top \\ \vdots \\ \boldsymbol{p}_{|\mathcal{P}|}^\top \end{bmatrix},$$

where each row corresponds to a point in the PCD. Analog to this notation, the normals are defined as $\mathcal{N} = \{\boldsymbol{n}_i\}_{i=1}^{|\mathcal{N}|} \subset \mathbb{R}^3$ with the corresponding matrix representation $\mathbf{N} \in \mathbb{R}^{|\mathcal{N}| \times 3}$. For a single tooth $\mathcal{P}_f$ the cardinalities $|\mathcal{P}_f| = |\mathcal{N}_f|$ are equal.

Due to the sequence-based nature of the model, a single object or patient is denoted as the set $\mathcal{S} = \{\mathcal{P}_f\}_{f\in\mathcal{F}}$. The number of teeth in the sequence is defined as $n_\mathrm{t} \coloneqq |\mathcal{F}| = |\mathcal{S}|$. Adjacent teeth are defined as the immediate neighbors of a target tooth as per the FDI numbering system, while antagonists are teeth from the opposing jaw within a fixed spatial vicinity of the target tooth. The sets of observed, adjacent and antagonist teeth are formally defined as $\mathcal{S}_\mathrm{obs}$, $\mathcal{S}_\mathrm{adj}$, $\mathcal{S}_\mathrm{anta}$, respectively, with the singleton target tooth $\mathcal{P}_\mathrm{tar}$.

2.1.3 Transformation representation

Per-tooth transformations are represented with respect to a global normalized patient dentition coordinate frame with translation, rotation and scale components. The set of per-tooth transforms is defined as $\mathcal{U} = \{\boldsymbol{u}_f\}_{f\in\mathcal{F}}$. Each tooth transform is represented as a vector

$$\boldsymbol{u} = [\boldsymbol{t}^\top, \boldsymbol{q}^\top, s]^\top \in \mathbb{R}^8,\ \boldsymbol{t} \in \mathbb{R}^3,\ \boldsymbol{q} = [q_w, q_x, q_y, q_z]^\top \in \mathbb{R}^4,\ s \in \mathbb{R} \tag{2.1}$$

with $\|\boldsymbol{q}\|_2 = 1$. Rotations are represented by unit quaternions $\boldsymbol{q}$ with the corresponding rotation matrix

$$\mathbf{R}(\boldsymbol{q}) = \begin{bmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y - q_wq_z) & 2(q_xq_z + q_wq_y) \\ 2(q_xq_y + q_wq_z) & 1-2(q_x^2+q_z^2) & 2(q_yq_z - q_wq_x) \\ 2(q_xq_z - q_wq_y) & 2(q_yq_z + q_wq_x) & 1-2(q_x^2+q_y^2) \end{bmatrix} \in \mathbb{R}^{3\times3}.$$

Conversely, for a proper rotation matrix $\mathbf{R}$ one closed form recovers $\boldsymbol{q}$ as

$$q_w = \tfrac{1}{2}\sqrt{1+\mathrm{tr}(\mathbf{R})},\quad q_x = \frac{R_{32}-R_{23}}{4q_w},\quad q_y = \frac{R_{13}-R_{31}}{4q_w},\quad q_z = \frac{R_{21}-R_{12}}{4q_w}, \tag{2.2}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace. Analogous expressions are used when $\mathrm{tr}(\mathbf{R}) \leq 0$ using the largest diagonal entry of $\mathbf{R}$. Intuitively, a unit quaternion $\boldsymbol{q} = (q_w, \mathbf{q}_v)$ encodes a 3D rotation by an angle θ around a unit axis $\mathbf{u}$ via $q_w = \cos(\theta/2)$ and $\mathbf{q}_v = \sin(\theta/2)\,\mathbf{u}$.

Quaternions provide a compact, numerically stable parameterization for 3D rotations with several advantages over rotation matrices and Euler angles:

- fewer parameters than matrices (4 vs. 9) with a simple unit-norm constraint,
- smooth interpolation with spherical linear interpolation,
- no gimbal lock or singularities as with Euler angles.

In this work, the main advantage is the compact representation of rotations with only four parameters, which in turn reduces the dimensionality of a full pose prediction task.

2.1.4 Operators

Several operators that are used throughout this work are defined. In this subsection, all symbols except the operator names are used locally for the purpose of defining the operators and do not imply meaning for identically named quantities elsewhere in this work.

Attention operators follow the notation of Vaswani et al. [60]. The standard scaled dot-product self attention is denoted as

$$\mathrm{Attn}(\mathbf{X}) = \mathrm{softmax}\left(\frac{Q(\mathbf{X})K(\mathbf{X})^\top}{\sqrt{d_k}}\right)V(\mathbf{X}), \tag{2.3}$$

where Q, K, and V are learned linear projections of the input $\mathbf{X}$, and d_k is the dimensionality of the key vectors. Similarly, $\mathrm{Attn}_{\mathrm{cross}}$ denotes the standard scaled dot-product cross-attention

$$\mathrm{Attn}_{\mathrm{cross}}(\mathbf{X}, \mathbf{Y}) = \mathrm{softmax}\left(\frac{Q(\mathbf{X})K(\mathbf{Y})^\top}{\sqrt{d_k}}\right)V(\mathbf{Y}). \tag{2.4}$$

The standard layer normalization [61] operator is defined for an input vector $\boldsymbol{x} \in \mathbb{R}^d$ as

$$\mathrm{LN}(\boldsymbol{x}) = \frac{\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]}{\sqrt{\mathrm{Var}(\boldsymbol{x}) + \epsilon}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}, \tag{2.5}$$

where $\mathbb{E}[\boldsymbol{x}]$ and $\mathrm{Var}(\boldsymbol{x})$ denote the mean and variance computed over the feature dimension, ϵ is a small constant for numerical stability, and $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^d$ are learnable affine parameters. The element wise Hadamard product is denoted by $\odot$.

The concat operator denotes concatenation along the feature dimension. For two input matrices $\mathbf{X} \in \mathbb{R}^{N \times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{N \times d_2}$, the concatenation is defined as

$$\mathrm{concat}(\mathbf{X}, \mathbf{Y}) = \begin{bmatrix} \mathbf{X} & \mathbf{Y} \end{bmatrix} \in \mathbb{R}^{N \times (d_1 + d_2)}.$$

The eig operator denotes the eigendecomposition of a square matrix. For a matrix $\mathbf{A} \in$

$\mathbb{R}^{n\times n}$ the eigendecomposition is given as

$$\text{eig}(\mathbf{A}) = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\top},$$

where $\mathbf{V} \in \mathbb{R}^{n\times n}$ is the matrix of eigenvectors and $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ is the diagonal matrix of eigenvalues.

2.2 Dataset and preprocessing

The dataset [62] used in this work comprises 1060 full-arch dental models in terminal occlusion sourced from 435 patients aged between 8 and 35 years that received orthodontic treatment. The complete dataset can include multiple scans per patient taken at different treatment stages without special annotations indicating these cases. Thus, every scan is treated as an independent patient instance in the following and referred to as a patient for simplicity.

The dataset is segmented and annotated according to the FDI tooth numbering system. Based on additional annotations, patients without permanent dentition and special orthodontic cases are excluded from the dataset. Furthermore, additional manual tooth-level annotations are performed to identify incomplete and damaged teeth. While these teeth are not excluded from the dataset, they are not used as restoration targets during training or evaluation. After filtering, the final dataset comprises 631 patients, each of which can be used for at least one of the two tooth ranges. For each patient and tooth, labels are generated that indicate whether a tooth can be used as a target tooth for reconstruction based on its integrity and completeness. Due to special adjacent tooth level orientation processing, teeth without two adjacent teeth are excluded from the potential target tooth set. The final dataset statistics are summarized in Table 2.1.

The dataset is split into training, validation and test sets at the patient level. Two separate splits are created for teeth in the range 1–3 (incisors and canines) and 4–6 (premolars and molars) including all quadrants. One sample corresponds to one target tooth with its observed context teeth.

Table 2.1: Dataset statistics after filtering for the two tooth ranges 1–3 and 4–6.

Mode	Range	Patients	Samples
train	1–3	570	5764
val	1–3	30	360
test	1–3	30	360
train	4–6	553	4610
val	4–6	30	360
test	4–6	30	360

2.2.1 Normalization and inverse mapping

To disentangle shape and pose variations across patients and teeth, while still preserving relative orientations to adjacent teeth, a normalization procedure is applied to all tooth PCDs in the dataset. The normalization consists of two steps.

First, patient-level normalization is performed to establish a canonical dentition coordinate frame. Second, tooth-level normalization is applied to center, rotate and scale each tooth within this global frame. The combined normalization ensures that all teeth are represented in a consistent coordinate frame that facilitates learning shape variations independently of pose differences. Notably, except for the tooth level translation of the target tooth, the normalization process is reversible due to reliance on observed data only. This reduces the TPP prediction space to translation only, reducing task complexity.

The final transformation operates on the tooth level using the transform $\boldsymbol{u} = [\boldsymbol{t}^\top, \boldsymbol{q}^\top, s]^\top$ according to (2.1) and (2.2). For a point $\boldsymbol{p} \in \mathbb{R}^3$ and normal $\boldsymbol{n} \in \mathbb{R}^3$ in patient space, the forward and inverse normalization operations are defined as

$$\text{forward (to normalized space):} \quad \boldsymbol{p}' = \frac{\mathbf{R}^\top \left(\boldsymbol{p} - \boldsymbol{t}\right)}{s}, \qquad \boldsymbol{n}' = \mathbf{R}\,\boldsymbol{n}, \tag{2.6}$$

$$\text{inverse (to patient space):} \quad \boldsymbol{p} = \mathbf{R}\left(s\,\boldsymbol{p}'\right) + \boldsymbol{t}, \qquad \boldsymbol{n} = \mathbf{R}\,\boldsymbol{n}'. \tag{2.7}$$

2.2.2 Patient-level normalization

To ensure a canonical global, patient-level base coordinate frame, a single reference transformation is computed per patient. A principal component analysis (PCA) of the full dentition, excluding gingival geometry, yields a right-handed coordinate frame centered

at the overall dental arch centroid. The PCDs of all teeth are concatenated along the point dimension to form $\mathbf{P}_{\text{pat}} = [\mathbf{P}_1^\top, \ldots, \mathbf{P}_{n_t}^\top]^\top \in \mathbb{R}^{(n_t|\mathcal{P}_f|)\times 3}$ and form the centered PCD

$$\mathbf{X}_{\text{pat}} = \mathbf{P}_{\text{pat}} - \mathbb{1}_{\text{pat}}\, \boldsymbol{t}_{\text{pat}}^\top, \qquad \boldsymbol{t}_{\text{pat}} = \bar{\mathbf{P}}_{\text{pat}}^\top \in \mathbb{R}^3, \qquad \mathbb{1}_{\text{pat}} \in \mathbb{R}^{n_t|\mathcal{P}_f|}. \tag{2.8}$$

The covariance is

$$\boldsymbol{\Sigma}_{\text{pat}} = \frac{1}{n_t|\mathcal{P}_f|}\, \mathbf{X}_{\text{pat}}^\top \mathbf{X}_{\text{pat}} \in \mathbb{R}^{3\times 3} \tag{2.9}$$

with its eigendecomposition

$$\text{eig}(\boldsymbol{\Sigma}_{\text{pat}}) = \mathbf{V}_{\text{pat}}\, \boldsymbol{\Lambda}_{\text{pat}}\, \mathbf{V}_{\text{pat}}^\top, \qquad \mathbf{V}_{\text{pat}} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3].$$

The eigenvectors $\mathbf{v}_i$ are initially sorted by decreasing eigenvalues λ_i to form an orthonormal frame. A right-handed basis ($\det \mathbf{V}_{\text{pat}} = +1$) is ensured by flipping $\mathbf{v}_3$, if necessary. The eigenvector matrix is then restructured such that the first two principal components are aligned with the occlusal plane that spans as a 2D barrier between the upper and lower jaws. The axis $\mathbf{v}_2$ is oriented from the posterior toward the anterior teeth and $\mathbf{v}_3$ points from the lower towards the upper jaw. The rotation matrix $\mathbf{R}_{\text{pat}} = \mathbf{V}_{\text{pat}}$ aligns the PCA frame with the standard basis according to (2.6) with

$$\mathbf{P}'_{\text{pat}} = \mathbf{R}_{\text{pat}}^\top (\mathbf{P}'_{\text{pat}} - \mathbb{1}_{\text{pat}} \boldsymbol{t}_{\text{pat}}^\top). \tag{2.10}$$

2.2.3 Tooth-level normalization

Normalization at the tooth level further refines the per-tooth transforms to ensure consistent orientation, centering, and scaling across all teeth in the dataset. This process involves quadrant-based orientation adjustments, centroid-based translations, PCA-based scaling, and adjacent tooth-based alignment refinements. It is to be noted that rotations to the normalized frame are defined as the transposed rotation matrices. For consistent orientation of teeth across jaws, teeth from the upper jaw (FDI quadrants $q \in \{1, 2\}$) are rotated to share the same orientation as the lower jaw using the fixed rotation

$$\mathbf{R}_{\text{quad}}^\top(q) = \begin{cases} \text{diag}(-1, 1, -1), & q \in \{1, 2\}, \\ \mathbb{I}_3, & q \in \{3, 4\}. \end{cases} \tag{2.11}$$

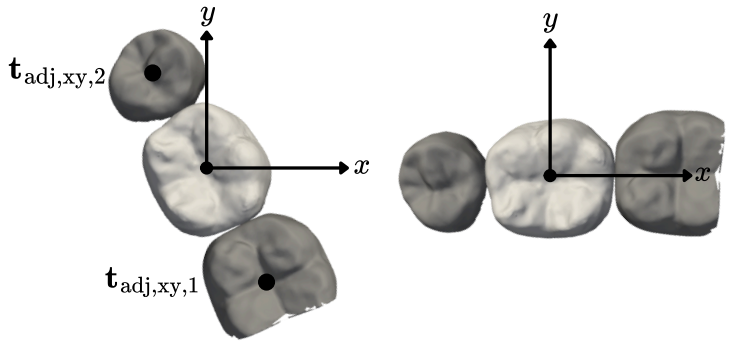
Thus, the relative position of the occlusal surfaces is consistent across jaws. Each tooth is then centered by its centroid $\boldsymbol{t}_f$ according to (2.8).

Adjacent-based orientation refinement uses the centroids of the two adjacent teeth per normalization target position in the patient-normalized frame (Fig. 2.2). Let $\boldsymbol{t}_{\text{adj,xy,1}} \in \mathbb{R}^2$, $\boldsymbol{t}_{\text{adj,xy,2}} \in \mathbb{R}^2$ denote the (x, y) components of the adjacent tooth centroids, where the ordering is quadrant- and FDI-dependent such that the resulting y-axis points outward (buccal, labial) relative to the dental arch. The adjacent direction vector is defined as $\boldsymbol{d} = \boldsymbol{t}_{\text{adj,xy,1}} - \boldsymbol{t}_{\text{adj,xy,2}} \in \mathbb{R}^2$. With the canonical outward reference $\mathbf{e}_y = (0, 1)^\top$, the refinement angle is given by $\alpha = \operatorname{atan2}\left((\boldsymbol{d})_y, (\boldsymbol{d})_x\right) - \operatorname{atan2}(1, 0)$, i.e., the signed in-plane angle between $\mathbf{e}_y$ and $\boldsymbol{d}$. The corresponding in-plane refinement is a pure rotation about the z-axis,

$$\mathbf{R}_{\text{adj}}^\top = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The dataset-level rotation used for tooth normalization then combines the jaw rotation $\mathbf{R}_{\text{quad}}(q)$ from (2.11) with the adjacent-based refinement as

$$\mathbf{R}_f^\top := \mathbf{R}_{\text{adj}}^\top \mathbf{R}_{\text{quad}}^\top(q).$$



(a) (b)

Fig. 2.2: Adjacent teeth-based tooth orientation refinement. (a) Teeth are aligned according to their position in the dental arch. (b) The proposed adjacent-based refinement aligns teeth according to the centroids of their immediate neighbors.

To obtain a reversible scaling, each tooth is temporarily displayed in its PCA frame $\mathbf{P}'_f$ following (2.9)-(2.10). The single tooth scale is defined as the mean extent of the PCD over all three axes in this frame. A patient-level scale is then computed as the average

over all observed teeth

$$s_{\text{pat}} = \frac{1}{|\mathcal{F}_{\text{obs}}|} \sum_{f \in \mathcal{F}_{\text{obs}}} s_f.$$

The per-tooth normalization is finally given by (2.6) with $\boldsymbol{t}_f$, $\mathbf{R}_f$, and s_{pat}. Furthest point sampling (FPS) [46] is applied to downsample each tooth PCD to a fixed number of points $|\mathcal{P}_{f,\text{fps}}|$ after normalization.

2.2.4 Antagonist selection

For functional crown design, antagonist teeth in the opposing jaw are essential to establish occlusal contacts. Contact analysis of two PCDs requires at least one point set to contain its normals. To avoid additional generative complexity, the reconstruction model is designed to predict target points without normals. Therefore, antagonist normals must be made available during training and inference. To focus on relevant regions for occlusion with minimal memory footprint, antagonist teeth are cropped to points within a spatial vicinity threshold ϑ_{anta} of the target tooth $\mathcal{P}_{\text{tar}}$. Let $\mathcal{P}_{\text{anta}}$ denote the PCD of the full antagonist jaw. The cropped antagonist point set is defined as

$$\mathcal{P}_{\text{anta,crop}} = \left\{ \boldsymbol{p}_{\text{anta}} \in \mathcal{P}_{\text{anta}} \mid \min_{\boldsymbol{p}_{\text{tar}} \in \mathcal{P}_{\text{tar}}} \|\boldsymbol{p}_{\text{anta}} - \boldsymbol{p}_{\text{tar}}\|_2 < \vartheta_{\text{anta}} \right\}. \tag{2.12}$$

The normals $\mathcal{N}_{\text{anta,crop}}$ are cropped according to the selected points in $\mathcal{P}_{\text{anta,crop}}$.

2.2.5 Curvature

To implement a curvature-weighted loss during TST training, per-point curvature values $\boldsymbol{\kappa} \in \mathbb{R}^{|\mathcal{P}_{\text{tar}}|}$ are computed on the target surface following the discrete curvature measure of [63]. The method starts from the observation that, on a triangulated surface, mean curvature is concentrated along edges and depends on the signed dihedral angle between the two incident triangle normals. Convex edges contribute positively to the integrated mean curvature, while concave edges contribute negatively. The resulting edge-based curvature contributions are then distributed to the adjacent vertices to obtain vertex-wise curvature values for training. In this work, signed curvatures are not required, and thus the absolute mean curvature values are used.

To obtain a local curvature estimate at a controllable spatial scale, curvature is evaluated within spherical neighborhoods of radius r_{curv} around query points. For each neighborhood, the curvature contributions of all mesh edges intersecting the ball are accumulated and then normalized by the corresponding surface patch area, yielding an estimate of the average mean curvature in that region. The radius r_{curv} determines the scale of the weighting. As shown in Fig. 2.3, small radii emphasize fine surface details, while larger radii capture broader shape features. In practice, r_{curv} must be chosen relative to the mesh resolution. If it is too small, neighborhoods may contain too few edges to produce meaningful variation, whereas overly large radii smooth out local curvature changes. Based on this trade-off, an intermediate radius of $r_{\text{curv}} = 0.8\,\text{mm}$ is used for all experiments in this work.

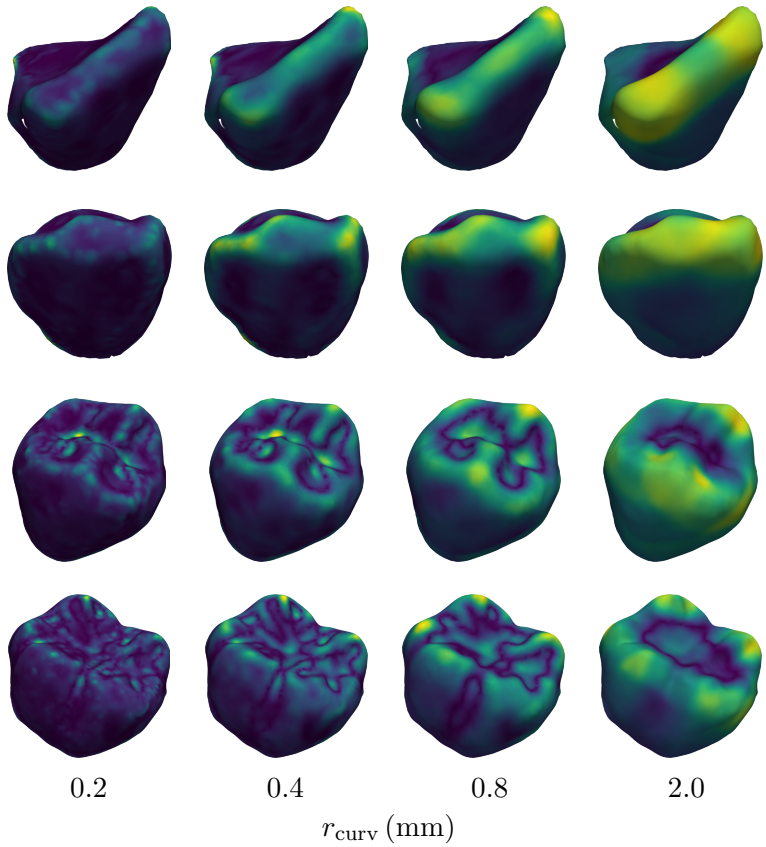


Fig. 2.3: Different curvature weighting radii r_{curv} for an exemplary patient and teeth with FDI codes 41, 43, 45, 46 from top to bottom. The colormaps are individually normalized for every tooth to highlight local curvature variations for each setting. A green shift in the colormap indicates higher curvatures.

2.3 Tooth Sequence Transformer

The following describes the reconstruction architecture, which maps the observed multi-tooth context to a dense target point set of fixed cardinality. All inputs are represented in the normalized space defined in Section 2.2.1.

2.3.1 Morphological and anatomical tooth encoder

Each observed tooth point set $\mathcal{P}_{\text{obs},f}$ is encoded by the point-set encoder

$$\phi_{\text{E,pts}} : \mathbb{R}^{|\mathcal{P}_{\text{obs},f}|\times 3} \to \mathbb{R}^C$$

into a geometric feature vector of width C. In parallel, each tooth position in the complete FDI sequence is associated with a learned identity embedding indexed by its FDI code via

$$\psi_{\text{fdi}} : f \to \mathbb{R}^C. \tag{2.13}$$

This identity embedding is constructed as the sum of a learned tooth embedding and a learned quadrant embedding, jointly capturing anatomical priors tied to local tooth type and global arch location. Identity embeddings are defined for all FDI positions irrespective of whether the corresponding tooth is present in the observed context.

To couple morphological observations with anatomical priors, the geometric feature and the identity embedding are concatenated and mapped back to the initial width through a fusion projection

$$\varphi : \mathbb{R}^{2\times C} \to \mathbb{R}^C. \tag{2.14}$$

Without additional conditioning, missing teeth and the reconstruction target are indistinguishable to the model because both lack an observed PCD. Dedicated role embeddings are therefore introduced and added on top of the FDI-based identity embeddings to explicitly encode the contextual function of each tooth position. Four learned role vectors are used,

$$\boldsymbol{e}_{\text{adj}}, \quad \boldsymbol{e}_{\text{anta}}, \quad \boldsymbol{e}_{\text{miss}}, \quad \boldsymbol{e}_{\text{tar}} \in \mathbb{R}^C. \tag{2.15}$$

These embeddings correspond to adjacent, antagonist, missing, and target teeth. They are shared across all teeth and patients and are optimized jointly with the remaining network parameters.

Following the residual-stream convention in transformer architectures, the per-tooth feature is formed by adding the appropriate role embedding to the fused morphological and anatomical encoding prior to any contextual aggregation,

$$\boldsymbol{h}_{\mathrm{obs},f} = \varphi\Big(\mathrm{concat}\Big(\phi_{\mathrm{E,pts}}\big(\mathcal{P}_{\mathrm{obs},f}\big), \psi_{\mathrm{fdi}}\big(f\big)\Big)\Big) + \boldsymbol{e}_{\mathrm{role}(f)}, \tag{2.16}$$

where $\mathrm{role}(f) \in \{\mathrm{adj}, \mathrm{anta}, \mathrm{miss}, \mathrm{tar}\}$ is determined by the tooth's relationship to the reconstruction target. Stacking all observed-tooth features yields the context matrix $\mathbf{H}_{\mathrm{obs}} \in \mathbb{R}^{|\mathcal{S}_{\mathrm{obs}}| \times C}$.

The target tooth feature is defined analogously, combining its anatomical identity embedding from (2.13) with the target role embedding from (2.15),

$$\boldsymbol{h}_{\mathrm{tar}} = \psi_{\mathrm{fdi}}\big(f_{\mathrm{tar}}\big) + \boldsymbol{e}_{\mathrm{tar}}, \tag{2.17}$$

thereby injecting both anatomical identity and target designation into the residual stream. The full context feature matrix over the complete tooth set $\mathcal{S}$, with $\mathcal{S}_{\mathrm{obs}} \subset \mathcal{S}$ and $\mathcal{P}_{\mathrm{tar}} \in \mathcal{S}$, is then given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\mathrm{obs}} \\ \boldsymbol{h}_{\mathrm{tar}}^{\top} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}| \times C}. \tag{2.18}$$

2.3.2 Contact Point Module

Contact points: Contact points are derived from occlusal interpenetrations between a tooth $\mathcal{P}_f$ and its antagonists $\mathcal{P}_{\mathrm{anta}}$ with normals $\mathcal{N}_{\mathrm{anta}}$ [64]. Nearest-neighbor correspondences between the tooth and the antagonist set provide point-to-surface distances $\mathcal{D}_{f,\mathrm{anta}}$ together with the corresponding antagonist normals $\mathcal{N}_{\mathrm{anta,nn}}$. These quantities are used to determine whether a tooth point lies in an occluded configuration relative to the opposing virtual antagonist surface.

Occlusion is decided via a signed contact-orientation indicator, defined as

$$\nu(\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{n}_2) = \frac{(\boldsymbol{p}_1 - \boldsymbol{p}_2)^{\top}}{\|\boldsymbol{p}_1 - \boldsymbol{p}_2\|_2} \boldsymbol{n}_2 \in [-1, 1]. \tag{2.19}$$

For each tooth point $\boldsymbol{p}_{f,i}$, the corresponding nearest antagonist point $\boldsymbol{p}_{\mathrm{anta,nn}(i)}$ and its normal $\boldsymbol{n}_{\mathrm{anta,nn}(i)}$ induce a signed occlusion distance by flipping the sign of the nearest-neighbor distance whenever the indicator signals an opposing orientation. This yields the

set of signed occlusion distances

$$\mathcal{D}_{\text{occl}} = \left\{ \begin{cases} -d_{f,\text{anta},i}, & \text{if } \nu\Big(\boldsymbol{p}_{f,i}, \boldsymbol{p}_{\text{anta,nn}(i)}, \boldsymbol{n}_{\text{anta,nn}(i)}\Big) < 0 \\ d_{f,\text{anta},i}, & \text{otherwise} \end{cases} \;\middle|\; i \in \{1, \ldots, |\mathcal{P}_f|\} \right\}. \quad (2.20)$$

Using the threshold d_{thresh}, tooth points are classified as occluded whenever the signed distance falls below the threshold,

$$\mathcal{P}_{\text{occl},f} = \{\boldsymbol{p}_{f,i} \mid d_{\text{occl},i} < d_{\text{thresh}},\, i \in \{1, \ldots, |\mathcal{P}_f|\}\}\,.$$

Contact features: For each observed tooth, occluded points are computed against the cropped antagonist set from (2.12) and reduced to a fixed cardinality contact set $\mathcal{P}_{\text{cp}}$ using FPS. If fewer occluded points than required are available, the remaining samples are filled by drawing additional points from subsequent nearest-neighbor correspondences in the antagonist set.

A shared contact encoder maps each tooth-specific contact set to a compact embedding. Concretely, an MLP followed by max-pooling aggregation [46] yields

$$\xi_{\text{cp}} : \mathbb{R}^{|\mathcal{P}_{\text{cp}}| \times 3} \to \mathbb{R}^C.$$

Teeth without detected contacts are assigned a learned no-contact embedding. Stacking over all observed teeth forms the contact feature matrix

$$\mathbf{H}_{\text{cp}} \in \mathbb{R}^{|\mathcal{S}_{\text{obs}}| \times C}.$$

Finally, contact information is injected into the geometric and anatomical tooth features via cross attention (2.4),

$$\mathbf{H}^{(1)} = \mathbf{H} + \text{Attn}_{\text{cross}}\Big(\mathbf{H}, \mathbf{H}_{\text{cp}}\Big).$$

The complete tooth sequence feature enrichment process is illustrated in Fig. 2.4.

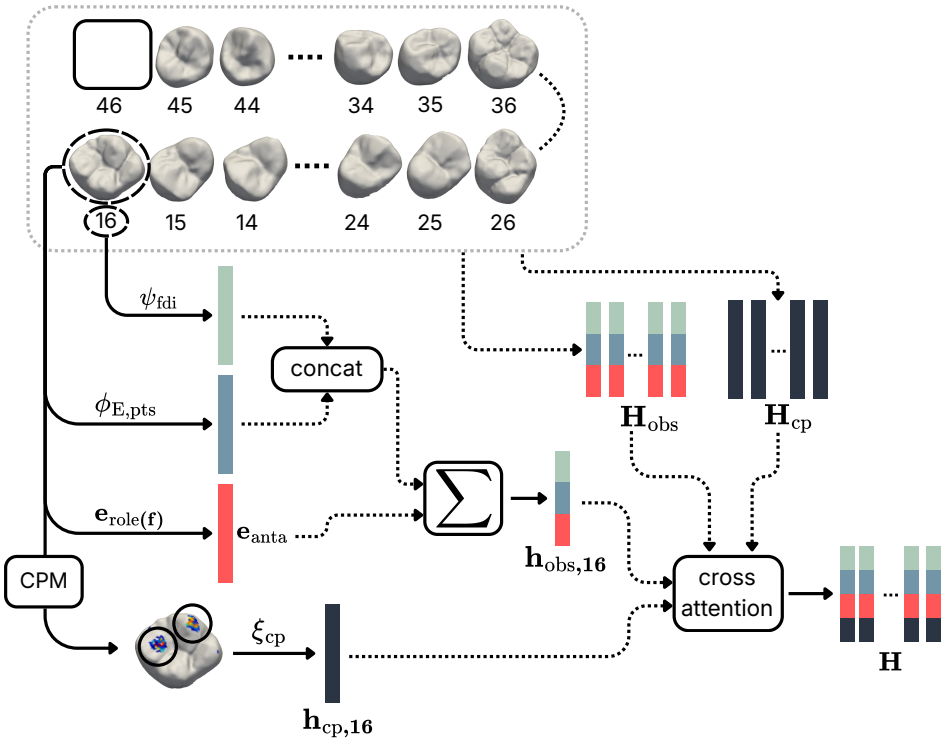


Fig. 2.4: Tooth sequence encoder with integrated CPM. Each observed tooth is encoded into a geometric feature and fused with an anatomical identity embedding derived from its FDI code. A role embedding is added to distinguish adjacent, antagonist, missing, and target teeth. Contact points are extracted via occlusion analysis against the antagonist set and encoded into compact contact features. These features are injected into the tooth sequence via cross attention, enriching the tooth-context representation with occlusal contact information prior to transformer encoding.

2.3.3 Transformer encoder

The transformer encoder is implemented as a stack of blocks operating on the stacked tooth features. Each block follows a pre-normalization residual design, applying self-attention and a subsequent MLP update to progressively enrich the tooth-context representation:

$$\mathbf{H}^{(1)} = \mathbf{H} + \text{Attn}\Big(\text{LN}(\mathbf{H})\Big),$$
$$\mathbf{H}^{(2)} = \mathbf{H}^{(1)} + \text{MLP}\Big(\text{LN}(\mathbf{H}^{(1)})\Big).$$

The attention and layer normalization operators are defined in (2.3) and (2.5), respectively.

Accordingly, each encoder block updates the stacked feature matrix via the mapping

$$\phi_{\mathrm{E,blk}} : \mathbb{R}^{|\mathcal{S}|\times C} \to \mathbb{R}^{|\mathcal{S}|\times C}, \qquad \phi_{\mathrm{E,blk}}(\mathbf{H}) = \mathbf{H}^{(2)}, \tag{2.21}$$

where $\phi_{\mathrm{E,blk}}$ denotes a single transformer block. Stacking multiple blocks yields the full transformer encoder.

In summary, each block exchanges information across the tooth sequence through self-attention and applies nonlinear refinement through the MLP. Contact information $\mathbf{H}_{\mathrm{cp}}$ is integrated via cross attention in a preceding step, ensuring that occlusal contact signals are incorporated into the tooth-context representation before being propagated and mixed across all encoder layers.

2.3.4 Coarse point generator

To decouple the encoder embedding width from the generator design during hyperparameter sweeps, $\boldsymbol{h}_{\mathrm{tar}}$ is first mapped to a fixed hidden dimension C_{H}. This design keeps the capacity and parameterization of the coarse generator constant while allowing the upstream transformer to vary in width.

The CPG is implemented as a learned projection

$$\rho : \mathbb{R}^{C_{\mathrm{H}}} \to \mathbb{R}^{Q\times 3}, \tag{2.22}$$

which transforms the hidden features through an MLP and outputs Q coarse 3D points. These points provide a low-resolution geometric scaffold that subsequent refinement stages can densify and locally adjust.

The resulting coarse point set $\mathcal{P}_{\mathrm{pred,c}}$ acts as a sparse representation of the target tooth, capturing global shape characteristics in a low-dimensional Euclidean space. Since the dense reconstruction is produced by a subsequent progressive translation-based expansion stage, the coarse points are not a subset of the final dense prediction $\mathcal{P}_{\mathrm{pred}}$. Instead, they serve as geometric anchors for constructing decoder queries.

To generate the Q geometry-informed query features that drive the transformer decoder, the coarse points are concatenated with the enriched and reshaped target feature $\boldsymbol{h}'_{\mathrm{tar,H}} \in$

$\mathbb{R}^{Q\times C_{\mathrm{H}}}$ and mapped to the decoder embedding width through a learned query projection

$$\Pi : \mathbb{R}^{(C_{\mathrm{H}}+3)} \to \mathbb{R}^{C}, \qquad \mathbf{Z} = \Pi\Big(\ \mathrm{concat}(\boldsymbol{h}'_{\mathrm{tar,H}}, \mathcal{P}_{\mathrm{pred,c}})\ \Big) \in \mathbb{R}^{Q\times C}.$$

Each query feature $\boldsymbol{z} \in \mathbb{R}^{C}$ associates a coarse geometric anchor with enriched target context, enabling the decoder to condition refinement on both global shape cues and the aggregated multi-tooth information.

2.3.5 Transformer decoder and Progressive Expansion Module

The final transformer decoder refines the geometry-informed target queries $\mathbf{Z}$ through (i) self-attention among query features and (ii) cross-attention to the encoded full tooth context $\mathbf{H} \in \mathbb{R}^{|\mathcal{S}|\times C}$. The decoder proceeds in two stages. First, an initial decoding stage updates the coarse query features while conditioning on the full tooth context. Second, multiple progressive expansion stages upsample the refined query features together with their associated points to obtain the final dense target prediction. Structurally, the decoder mirrors the encoder, but operates on query tokens that attend both within the query set and to the encoder feature matrix.

Stage 1: Feature refinement. Each decoder block comprises three sub-layers, namely self-attention, cross-attention, and a feed-forward MLP. Each sub-layer is applied in a residual configuration with layer normalization

$$\mathbf{Z}^{(1)} = \mathbf{Z} + \ \mathrm{Attn}\Big(\ \mathrm{LN}(\mathbf{Z})\Big), \tag{2.23}$$

$$\mathbf{Z}^{(2)} = \mathbf{Z}^{(1)} + \ \mathrm{Attn}_{\mathrm{cross}}\Big(\ \mathrm{LN}(\mathbf{Z}^{(1)}),\ \mathrm{LN}(\mathbf{H})\Big), \tag{2.24}$$

$$\mathbf{Z}^{(3)} = \mathbf{Z}^{(2)} + \mathrm{MLP}\Big(\ \mathrm{LN}(\mathbf{Z}^{(2)})\Big). \tag{2.25}$$

Consequently, each decoder block implements the mapping

$$\phi_{\mathrm{D,blk}} : \mathbb{R}^{Q\times C} \to \mathbb{R}^{Q\times C}, \qquad \phi_{\mathrm{D,blk}}(\mathbf{Z}, \mathbf{H}) = \mathbf{Z}^{(3)}, \tag{2.26}$$

where $\phi_{\mathrm{D,blk}}$ denotes a single decoder block. Stacking multiple blocks yields the complete transformer decoder, producing refined query features that are subsequently upsampled by the progressive expansion stages to obtain the dense target PCD.

After the first decoding stage, the coarse prediction $\mathcal{P}_{\text{pred,c}} \in \mathbb{R}^{Q\times 3}$ and its associated query features $\mathbf{Z} \in \mathbb{R}^{Q\times C}$ are progressively upsampled to the full target prediction $\mathcal{P}_{\text{pred}}$ through attention-driven feature refinement [65] and point expansion within the PEM. Motivated by prior work on PCD completion and upsampling, offset-based point expansion is applied in multiple stages to gradually increase point density [51, 52, 66]. Instead of regressing absolute coordinates for newly created points, the model iteratively predicts local offset vectors that are added to existing point locations. Direct coordinate regression tends to become unstable and noisy as the point count increases, whereas offset-based prediction encourages learning local geometric relationships that are easier to model and generalize [66].

The number of expansion stages is denoted by $n_{\text{pe}} \in \mathbb{N}$. For convenience, the query feature matrix and point sets are denoted by their set equivalents $\mathcal{Z}$ and $\mathcal{P}$ in the following. Each stage $i \in \{1, \dots, n_{\text{pe}}\}$ operates on a point set and its associated feature set,

$$\mathcal{P}^{(i)} \in \mathbb{R}^{n_{\text{pts},i}\times 3}, \qquad \mathcal{Z}^{(i)} \in \mathbb{R}^{n_{\text{pts},i}\times C},$$

initialized with $\mathcal{P}^{(0)} = \mathcal{P}_{\text{pred,c}}$ and $\mathcal{Z}^{(0)} = \mathcal{Z}$. Each stage then produces a denser point set $\mathcal{P}^{(i+1)}$ together with updated features $\mathcal{Z}^{(i+1)}$, which are passed to the next expansion stage.

Expansion schedule. Let the initial and final point counts be $n_{\text{pts},0} = Q = |\mathcal{P}_{\text{pred,c}}|$ and $n_{\text{pts},n_{\text{pe}}} = |\mathcal{P}_{\text{pred}}|$, respectively. Intermediate stage sizes are chosen to follow a geometric progression between the coarse and final resolutions. The per-stage expansion factor and the resulting stage-wise cardinalities are defined as

$$c_{\text{pe}} = \left(\frac{n_{\text{pts},n_{\text{pe}}}}{n_{\text{pts},0}}\right)^{1/n_{\text{pe}}}, \qquad n_{\text{pts},i} = n_{\text{pts},0} c_{\text{pe}}^{i}, \qquad n_{\text{pts},i} \in \mathbb{N}, \qquad i = 1, \dots, n_{\text{pe}},$$

with $n_{\text{pts},n_{\text{pe}}}$ fixed to the desired final cardinality of the target tooth prediction.

Stage 2: Feature refinement. Each expansion stage first updates the current per-point features through self-attention within the stage-specific point set and cross-attention to the encoded multi-tooth context $\mathbf{H}$. For stage i, the feature update follows the same decoder pattern described in (2.23)–(2.25), using the current point features $\mathcal{Z}^{(i)}$ as input. This refinement step enriches each point representation using interactions within the evolving

point set while preserving conditioning on global tooth context, thereby preparing the features for the subsequent generation of additional points.

Translation-based expansion. Let $\mathcal{Z} = \{\boldsymbol{z}_j\}_{j=1}^{n_{\text{pts},i}}$ and $\mathcal{P} = \{\boldsymbol{p}_j\}_{j=1}^{n_{\text{pts},i}}$. For each stage, per-point features and coordinates are concatenated row-wise and passed to a learned offset predictor,

$$\Delta_i : \mathbb{R}^{C+3} \to \mathbb{R}^{3c_{\text{pe},i}}, \tag{2.27}$$

which is implemented as an MLP. Applied to each point, the predicted offsets

$$\left[\, \boldsymbol{\delta}_{j,1}^{(i)\top} \;\cdots\; \boldsymbol{\delta}_{j,c_{\text{pe},i}}^{(i)\top} \,\right]^\top = \Delta_i\Big(\operatorname{concat}(\boldsymbol{z}_j^{(i-1)}, \boldsymbol{p}_j^{(i-1)})\Big)$$

define $c_{\text{pe},i}$ local translations around each reference location $\boldsymbol{p}_j^{(i-1)}$ with $\boldsymbol{\delta}_{j,\ell}^{(i)} \in \mathbb{R}^3$, $\ell = 1, \ldots, c_{\text{pe},i}$, and $j = 1, \ldots, n_{\text{pts},i}$. The expanded point set is then obtained by translating the reference points by the predicted offsets,

$$\boldsymbol{p}_{j,\ell}^{(i)} = \boldsymbol{p}_j^{(i-1)} + \boldsymbol{\delta}_{j,\ell}^{(i)}, \qquad j = 1, \ldots, n_{\text{pts},i},\ \ell = 1, \ldots, c_{\text{pe},i}, \tag{2.28}$$

and subsequently re-indexing into $\mathcal{P}^{(i)} \in \mathbb{R}^{n_{\text{pts},i} \times 3}$ with $n_{\text{pts},i} = n_{\text{pts},i-1} c_{\text{pe},i}$. Thus, each point at stage i spawns $c_{\text{pe},i}$ descendants in its local neighborhood, conditioned on both the stage-wise decoder features and the global tooth context, with $\mathcal{P}^{(i-1)} \not\subset \mathcal{P}^{(i)}$. In parallel, the associated features are expanded by applying a linear projection that replicates each $\boldsymbol{z}_j^{(i-1)}$ by $c_{\text{pe},i}$, yielding $\mathcal{Z}^{(i)} \in \mathbb{R}^{n_{\text{pts},i} \times C}$.

Iterating the feature refinement in (2.23)–(2.25) and the translation-based expansion in (2.27)–(2.28) over all n_{pe} stages produces the final dense prediction $\mathcal{P}_{\text{pred}} = \mathcal{P}^{(n_{\text{pe}})}$. The initial coarse prediction is not a subset of the resulting points. Each fine-scale point is realized as a learned translation of a coarse ancestor, conditioned on its local decoded representation and the global multi-tooth context.

2.3.6 Training objective

The generic CD [67] is a widely used measure for quantifying similarity between two PCDs. It aggregates nearest-neighbor distances in both directions, thereby penalizing missing regions and spurious predictions. Prior work indicates that emphasizing high-curvature regions during optimization can improve the reconstruction of anatomically

salient structures such as cusps and ridges [36]. Motivated by this observation, training employs a curvature-weighted chamfer loss:

$$\mathcal{L}_{\text{tst,ccd}}(\mathcal{P}_{\text{tar}}, \mathcal{P}_{\text{pred}}) = \frac{1}{|\mathcal{P}_{\text{tar}}|} \sum_{i=1}^{|\mathcal{P}_{\text{tar}}|} e^{\beta_{\text{ccd}}\, \tilde{\kappa}(\boldsymbol{p}_{\text{tar},i})} \min_j \|\boldsymbol{p}_{\text{tar},i} - \boldsymbol{p}_{\text{pred},j}\|_2 + \frac{1}{|\mathcal{P}_{\text{pred}}|} \sum_{j=1}^{|\mathcal{P}_{\text{pred}}|} e^{\beta_{\text{ccd}}\, \tilde{\kappa}(\boldsymbol{p}_{\text{tar},i})} \min_i \|\boldsymbol{p}_{\text{pred},j} - \boldsymbol{p}_{\text{tar},i}\|_2.$$

The hyperparameter β_{ccd} controls the influence of curvature weighting. Setting $\beta_{\text{ccd}} = 0$ recovers the standard unweighted chamfer loss. To make the weighting independent of scale and tooth type, curvature values $\boldsymbol{\kappa}$ are normalized to the range $[0, 1]$ using the infinity norm

$$\tilde{\kappa}_i = \frac{|\kappa_i|}{\|\boldsymbol{\kappa}\|_\infty}.$$

2.4 Tooth Pose Predictor

The TST framework represents tooth geometry and pose in a disentangled manner. A dedicated TPP is therefore used to recover the partially reversible target-to-patient transformation $\boldsymbol{u}_{\text{pred}}$. By design, the scale and rotation components are reversible. In contrast, the translation relative to the patient-level base frame cannot be recovered from normalized geometry alone and must be inferred from context. Accordingly, this stage predicts the target translation vector $\boldsymbol{t}_{\text{pred}} \in \mathbb{R}^3$ from the observed sequence of per-tooth translations $\mathcal{T}_{\text{obs}} = \{\boldsymbol{t}_{\text{obs},f}\}_{f \in \mathcal{F}_{\text{obs}}}$.

2.4.1 Adjacent-mean translation normalization

To improve prediction stability, the translation sequence is normalized with respect to the mean position of the target tooth's adjacent teeth. This normalization reduces patient-specific global offsets and encourages the network to learn adjacent-relative translation patterns. The adjacent mean translation is

$$\bar{\boldsymbol{t}}_{\text{adj}} = \frac{1}{|\mathcal{F}_{\text{adj}}|} \sum_{f \in \mathcal{F}_{\text{adj}}} \boldsymbol{t}_f, \qquad \boldsymbol{t}_f \in \mathbb{R}^3.$$

For each observed tooth, the normalized translation is then computed as

$$\boldsymbol{t}_{f,\text{norm}} = \boldsymbol{t}_f - \bar{\boldsymbol{t}}_{\text{adj}}$$

to obtain the set of normalized translations that serves as input to the network. The absolute target translation is recovered at the end of the forward pass.

2.4.2 Transformer architecture

Since the TPP operates on a substantially lower-dimensional input space than the TST, the morphological point encoder is replaced by a compact ResNet-style [68] embedding network,

$$\phi_{\text{E,tpp}} : \mathbb{R}^3 \to \mathbb{R}^{C'} \tag{2.29}$$

implemented as a residual MLP [69]. This encoder maps each translation vector into the C'-dimensional feature space.

Using these translation embeddings, the TPP reuses the anatomical embeddings in (2.13), the linear fusion in (2.14), the role embeddings in (2.15), and the residual-stream feature construction in (2.16), adapted to the pose prediction setting. This yields the stacked context feature matrix $\mathbf{H}_{\text{obs,tpp}}$.

The target tooth embedding $\boldsymbol{h}_{\text{tar,tpp}}$ is represented without geometric encoding according to (2.17). The full TPP feature matrix is therefore assembled as $\mathbf{H}_{\text{tpp}}$ following (2.18). Finally, cross-attention-based pooling uses $\boldsymbol{h}_{\text{tar,tpp}}$ as the query and $\mathbf{H}_{\text{obs,tpp}}$ as keys and values, yielding the target context feature $\boldsymbol{h}_{\text{attn,tpp}} \in \mathbb{R}^{C'}$. The aggregated feature thus captures multi-tooth translation patterns relevant to predicting the target translation.

The prediction head mirrors the residual MLP structure of (2.29) and is implemented as a decoder with the mapping

$$\omega : \mathbb{R}^{C'} \to \mathbb{R}^3.$$

It maps the aggregated context feature $\boldsymbol{h}_{\text{attn,tpp}}$ to a translation in the normalized adjacent frame

$$\boldsymbol{t}_{\text{pred,norm}} = \omega(\boldsymbol{h}_{\text{attn,tpp}}) \in \mathbb{R}^3$$

and the final absolute prediction is recovered by inverting the adjacent-mean shift

$$\boldsymbol{t}_{\text{pred}} = \boldsymbol{t}_{\text{pred,norm}} + \bar{\boldsymbol{t}}_{\text{adj}}.$$

2.4.3 Training objective

The target translation corresponds to the translation component of the target transform vector $\boldsymbol{u}_{\text{tar}}$. Training minimizes a mean-squared error between the predicted and target translations expressed in the patient-level base frame

$$\mathcal{L}_{\text{tpp}} = \left\| \boldsymbol{t}_{\text{pred}} - \boldsymbol{t}_{\text{tar}} \right\|_2^2. \tag{2.30}$$

2.5 Tooth Pose Optimizer

The final stage refines the pose of the reconstructed crown in the patient-specific global frame. Starting from the normalized reconstruction produced by the TST (Section 2.3) and the initial translation estimate provided by the TPP (Section 2.4), the TPO optimizes a rigid translation together with an isotropic scale. The objective is to obtain occlusal and proximal contacts that match prescribed targets while respecting minimal thickness constraints with respect to an optional preparation surface.

2.5.1 Mesh generation

The raw output of the TST stage is a dense PCD representation of the target tooth. Since downstream dental processing typically requires explicit surface representations, a surface reconstruction stage is integrated into the TPO inference pipeline.

Surface normals are estimated via kNN-based tangent plane fitting. For each point $\boldsymbol{p}_{\text{pred},i} \in \mathcal{P}_{\text{pred}}$, the normal vector $\boldsymbol{n}_{\text{pred},i}$ is computed on $\mathcal{P}_{\text{pred}}$ using a local kNN tangent plane with consistent orientation. To stabilize the global normal direction, an orientation reference $\boldsymbol{r}$ is set to the centroid of the antagonist tooth. The mean dot product

$$\frac{1}{|\mathcal{P}_{\text{pred}}|} \sum_{i=1}^{|\mathcal{P}_{\text{pred}}|} \langle \boldsymbol{n}_{\text{pred},i}, \frac{\boldsymbol{r}}{\|\boldsymbol{r}\|_2} \rangle$$

is then used to flip all normals, if they oppose the reference, promoting outward-facing normals for the subsequent reconstruction stage.

For surface reconstruction, NKSR [59] is employed, as it reliably preserves high-curvature anatomy that is prevalent in dental crowns, in particular within occlusal regions of poste-

rior teeth. Compared to classical methods such as Ball-Pivoting [70] and Poisson Surface Reconstruction [71], NKSR produces sharper anatomical detail while still yielding watertight meshes.

The normalized representation of the PCD further supports reconstruction by enforcing a consistent global orientation, ensuring proper alignment of normals with $+\boldsymbol{e}_z$ prior to meshing.

NKSR reconstructs a watertight surface from an oriented point cloud by first estimating a sparse, hierarchical voxel structure around the underlying shape using a sparse convolutional network. On this hierarchy, an implicit Neural Kernel Field is defined from local, compactly supported kernel functions and conditioned on learned voxel features. The kernel weights are computed by solving a sparse linear system that fits the input geometry while enforcing consistency with the predicted voxel normals. The implicit field is then evaluated at grid corners of the voxel hierarchy and polygonized by extracting the zero level set using dual marching cubes [59].

Fig. 2.5 shows example reconstructions from different methods starting from the same predicted TST point cloud. Compared to Ball-Pivoting and Poisson surface reconstruction, NKSR better preserves high-curvature features such as cusps and ridges and yields watertight meshes without extensive parameter tuning.

Fig. 2.5: Reconstructed surfaces obtained with different methods from the same predicted TST PCD containing 2048 points. For NKSR, the detail level is set to 0.95. For Ball-Pivoting, the ball radii are chosen as multiples of the average nearest-neighbor distance using factors $\{2, 4, 8, 16\}$. For Poisson surface reconstruction, the reconstruction depth is set to 9.

To motivate the choice of the TST output cardinality and, consequently, the NKSR input resolution, Fig. 2.6 compares reconstructions obtained from different numbers of input points sampled from the same ground truth (GT) mesh. Reconstructions from 2048 points already exhibit no visible artifacts and retain anatomical detail, providing a favorable trade-off between geometric fidelity and computational cost.

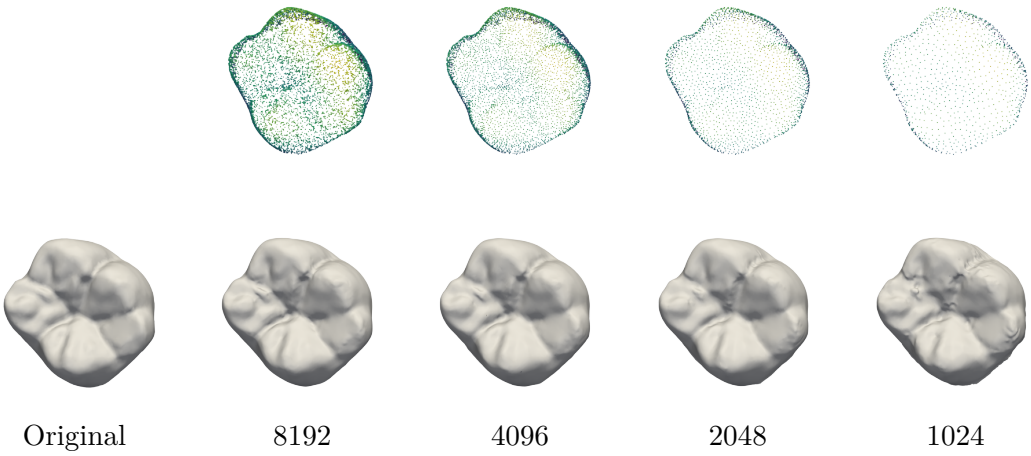


Fig. 2.6: NKSR reconstructions for different input point counts sampled from the same GT mesh. The NKSR detail level is set to 0.95 in all cases. Columns show the original mesh (left) and reconstructions from 8192, 4096, 2048, and 1024 points (right). The upper row visualizes the corresponding input points used for reconstruction.

Fig. 2.7 compares NKSR reconstructions obtained from different prediction networks at two reconstruction detail levels. For PoinTr and PCN, surface noise in the predicted PCDs leads to noticeably degraded mesh quality at the higher detail setting relative to TST. This illustrates that the NKSR detail level must be chosen in accordance with the noise characteristics of the input PCD to obtain stable and anatomically plausible reconstructions for these baselines.

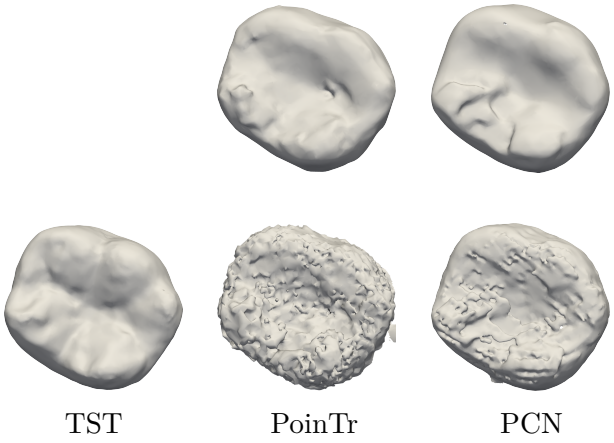


Fig. 2.7: NKSR reconstructed surfaces from different prediction networks. The detail level for NKSR is set to 0.95 (bottom row) and 0.5 (top row).

2.5.2 Optimization setup and initialization

The reconstructed mesh vertices provide a dense PCD representation that can be used directly for pose optimization. Let $\mathcal{P}_{\text{pred,vert}}$ denote the normalized mesh vertices and let $\boldsymbol{t}_{\text{pred}}$ be the translation predicted by the TPP. The partially reversible patient-level transformation is given by

$$\boldsymbol{u}_{\text{tar}} = \left[\left[0, 0, 0\right], \boldsymbol{q}^\top, s \right]^\top \in \mathbb{R}^8.$$

The normalized prediction $\mathcal{P}_{\text{pred,vert}}$ is partially inverted using $\boldsymbol{u}_{\text{tar}}$ according to (2.7), yielding the initial target state for optimization $\mathcal{P}_{\text{opt},0}$. Optimization is performed over an absolute translation $\boldsymbol{t}_{\text{opt}} = [t_x, t_y, t_z]^\top \in \mathbb{R}^3$ and an isotropic scale $s_{\text{opt}} \in \mathbb{R}$. The translation is initialized with $\boldsymbol{t}_{\text{pred}}^\top$ and the scale is initialized with 1.0. The target state at optimization step i is defined as

$$\mathcal{P}_{\text{opt},i} = \left\{ s_{\text{opt},i}\, \boldsymbol{p}_{\text{opt},0,j} + \boldsymbol{t}_{\text{opt},i} \mid \boldsymbol{p}_{\text{opt},0,j} \in \mathcal{P}_{\text{opt},0} \right\}, \qquad i = 1, \ldots, n_{\text{opt}}.$$

The objective is evaluated against the adjacent teeth $\mathcal{P}_{\text{adj},1}$ and $\mathcal{P}_{\text{adj},2}$ and the antagonist tooth set $\mathcal{P}_{\text{anta}}$. For proximal targets, the adjacent teeth are represented as separate objects, whereas the antagonists can be aggregated into a single PCD. Each reference PCD and its corresponding normals $\mathcal{N}$ are transformed into the patient-level base frame according to (2.7). If preparation data is available, the prepared tooth surface $\mathcal{P}_{\text{prep}}$ is provided as an additional PCD in the same frame together with its normals $\mathcal{N}_{\text{prep}}$.

After optimization, a final rotation and translation with $\mathbf{R}_{\text{pat}}$ and $\boldsymbol{t}_{\text{pat}}$ maps the optimized crown back to the world frame, yielding $\mathcal{P}_{\text{opt},n_{\text{opt}},\text{world}}$.

2.5.3 Contact evaluation

The penetration-vicinity function quantifies both interpenetration and near-contact configurations between two surfaces. It operates on the nearest-neighbor distance set $\mathcal{D}_{\text{nn}}$ between two PCDs and classifies each correspondence using the orientation indicator function $\nu(\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{n}_2)$ defined in (2.19).

Penetration and vicinity values are initialized as:

$$p_{\text{init}} = \frac{1}{|\mathcal{D}_{\text{pen}}|} \sum_{i=1}^{|\mathcal{D}_{\text{pen}}|} d_{\text{pen},i}, \qquad \mathcal{D}_{\text{pen}} = \left\{ d_{\text{nn},i} \mid \nu(\boldsymbol{p}_{1,i}, \boldsymbol{p}_{2,\text{nn}(i)}, \boldsymbol{n}_{2,\text{nn}(i)}) < 0 \right\}_{i=1}^{|\mathcal{P}_1|},$$
$$v_{\text{init}} = \min\left(\mathcal{D}_{\text{vic}}\right), \qquad \mathcal{D}_{\text{vic,init}} = \left\{ d_{\text{nn},i} \mid \nu(\boldsymbol{p}_{1,i}, \boldsymbol{p}_{2,\text{nn}(i)}, \boldsymbol{n}_{2,\text{nn}(i)}) > 0 \right\}_{i=1}^{|\mathcal{P}_1|}.$$

To evaluate penetration and vicinity separately, a sequential mutual-gating mechanism is applied. Vicinity is only evaluated if no penetration is detected. Given a gating threshold ϑ_{gate}, the final metrics are

$$p = \begin{cases} p_{\text{init}}, & v_{\text{init}} < \vartheta_{\text{gate}}, \\ 0, & \text{otherwise}, \end{cases} \qquad v = \begin{cases} v_{\text{init}}, & p < \vartheta_{\text{gate}}, \\ 0, & \text{otherwise}. \end{cases} \tag{2.31}$$

2.5.4 Optimization objective

To construct the optimization objective, the penetration and vicinity metrics in (2.31) are evaluated between the optimized target point set $\mathcal{P}_{\text{opt},i}$ and each relevant structure. Namely, the adjacent teeth $\mathcal{P}_{\text{adj},j}$ with $j \in \{1, 2\}$, the antagonist set $\mathcal{P}_{\text{anta}}$, and the prepared tooth surface $\mathcal{P}_{\text{prep}}$ are considered. This yields per-object scalars

$$p_k, \quad v_k, \qquad k \in \{(\text{adj}, 1), (\text{adj}, 2), (\text{anta}), (\text{prep})\},$$

which quantify the individual penetration and vicinity contributions that enter the overall objective.

Target occlusal contact values are derived once per case by evaluating $\mathcal{L}_{\text{pen}}$ and $\mathcal{L}_{\text{vic}}$ between the adjacent teeth and the antagonist, yielding $\tilde{p}_{\text{anta}}$ and $\tilde{v}_{\text{anta}}$. Target values for proximal contacts with the adjacent teeth are specified manually as $\tilde{p}_{\text{adj}}$ and $\tilde{v}_{\text{adj}}$. If preparation data is available, the vicinity target for the preparation is set to the minimal thickness threshold $\tilde{v}_{\text{prep}} = \vartheta_{\text{prep}}$, with zero penetration permitted.

The objective comprises four terms designed to promote symmetric proximal contacts, enforce minimal thickness with respect to the preparation, and match a desired occlusal contact profile with the antagonists.

The adjacency equality term encourages symmetric proximal contact conditions

$$\mathcal{L}_{\text{eq}} = |p_{\text{adj},1} - p_{\text{adj},2}| + |v_{\text{adj},1} - v_{\text{adj},2}|.$$

The adjacency absolute term matches penetration and vicinity to the prescribed adjacent contact targets

$$\mathcal{L}_{\text{adj}} = \sum_{j=1}^{2} \Big(|p_{\text{adj},j} - \tilde{p}_{\text{adj}}| + |v_{\text{adj},j} - \tilde{v}_{\text{adj}}| \Big). \tag{2.32}$$

The antagonist term enforces the desired occlusal contact profile, which can be specified manually or derived from the adjacent–antagonist configuration:

$$\mathcal{L}_{\text{anta}} = |p_{\text{anta}} - \tilde{p}_{\text{anta}}| + |v_{\text{anta}} - \tilde{v}_{\text{anta}}|.$$

Finally, when applicable, the preparation term enforces the minimal thickness constraint while preventing penetration into the prepared surface

$$\mathcal{L}_{\text{prep}} = |v_{\text{prep}} - \tilde{v}_{\text{prep}}| + p_{\text{prep}}.$$

The overall TPO objective is defined as a weighted sum of these four components:

$$\mathcal{L}_{\text{tpo}} = \lambda_{\text{eq}} \mathcal{L}_{\text{eq}} + \lambda_{\text{adj}} \mathcal{L}_{\text{adj}} + \lambda_{\text{anta}} \mathcal{L}_{\text{anta}} + \lambda_{\text{prep}} \mathcal{L}_{\text{prep}}. \tag{2.33}$$

Depending on the clinical scenario, user preference or the tooth type, terms in the objective can be selectively activated or deactivated by setting the corresponding weights to zero.

2.6 Metrics

Quantitative evaluation in dental 3D reconstruction commonly draws on general-purpose 3D shape metrics, and the specific metric choices and conventions vary across the literature [15]. Following the protocol of Broll et al. [12, 64], the evaluation employs a combined set of morphological and contact-based metrics to assess reconstruction quality (Table 2.2).

Table 2.2: Evaluation metrics. The table is adapted from Broll et al. [12].

Metric	Unit	Purpose
$\mathcal{L}_{\text{cd}}$	mm	General morphological accuracy.
$\mathcal{L}_{\text{cIoU}}$	–	Spatial alignment and fit in the tooth gap.
$\mathcal{L}_{\text{pen}}$	mm	Occlusal contact strength.
$\mathcal{L}_{\text{cp,pos}}$	mm	Shape of the occlusal contact point pattern.
$\mathcal{L}_{\text{cp,dist}}$	mm	Spatial spread of the occlusal contact points.
$\mathcal{L}_{\text{cp,num}}$	–	Number of occlusal contact points.
$\mathcal{L}_{\text{adj}}$	mm	Proximal contact penetration depth.

2.6.1 Morphological metrics

Morphological similarity is quantified using the mean two-sided CD with non-squared $\mathcal{L}_2$ distances

$$\mathcal{L}_{\text{cd}}(\mathcal{P}_{\text{pred}}, \mathcal{P}_{\text{GT}}) = \frac{1}{2}\left(\frac{1}{|\mathcal{P}_{\text{pred}}|}\sum_{i=1}^{|\mathcal{P}_{\text{pred}}|}\min_j \|\boldsymbol{p}_{\text{pred},i} - \boldsymbol{p}_{\text{GT},j}\|_2 + \frac{1}{|\mathcal{P}_{\text{GT}}|}\sum_{j=1}^{|\mathcal{P}_{\text{GT}}|}\min_i \|\boldsymbol{p}_{\text{GT},j} - \boldsymbol{p}_{\text{pred},i}\|_2\right),$$

and the complemented intersection over union between the axis-aligned bounding boxes of the reconstructed target tooth PCD and the GT

$$\mathcal{L}_{\text{cIoU}}(\mathcal{P}_{\text{pred}}, \mathcal{P}_{\text{GT}}) \approx \mathcal{L}_{\text{cIoU}}(\mathcal{P}_{\text{pred,bb}}, \mathcal{P}_{\text{GT,bb}}) = 1 - \frac{|\mathcal{P}_{\text{pred,bb}} \cap \mathcal{P}_{\text{GT,bb}}|}{|\mathcal{P}_{\text{pred,bb}} \cup \mathcal{P}_{\text{GT,bb}}|}.$$

2.6.2 Contact metrics

Occlusion-specific metrics: Occlusion-specific metrics quantify the contact situation between the reconstructed tooth and its antagonists relative to the GT contact pattern. Occluded points $\mathcal{P}_{\text{occl}}$ are computed according to (2.20). Contact points are then defined as the mean-shift cluster centers M of $\mathcal{P}_{\text{occl}}$ [64, 72].

Penetration is quantified by the set of occlusal distances below the penetration threshold between the reconstructed occlusal surface and the antagonists

$$\mathcal{D}_{\text{pen}} = \{d_{\text{occl},i} \mid d_{\text{occl},i} < d_{\text{thresh}},\, i \in \{1, \ldots, |\mathcal{D}_{\text{occl}}|\}\}\,.$$

The loss compares the mean penetration depth of the GT and the prediction with

$$\mathcal{L}_{\text{pen}} = \sqrt{\left(\frac{1}{|\mathcal{P}_{\text{pen,gt}}|} \sum_{i=1}^{|\mathcal{P}_{\text{pen,gt}}|} d_{\text{pen,gt},i} - \frac{1}{|\mathcal{P}_{\text{pen,pred}}|} \sum_{i=1}^{|\mathcal{P}_{\text{pen,pred}}|} d_{\text{pen,pred},i} \right)^2}.$$

To quantify the spatial spread of contact points, pairwise Euclidean distances between contact-point centers $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ with $i, j \in \{1, \ldots, |M|\}$ are evaluated

$$d_{ij} = \|\boldsymbol{m}_i - \boldsymbol{m}_j\|_2,$$

forming the distance matrix $\boldsymbol{D}$. The mean off-diagonal distance

$$d_{\text{od}} = \frac{1}{|M|(|M|-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{|M|} d_{ij}$$

is used to compare GT and predicted contact patterns via

$$\mathcal{L}_{\text{cp,dist}} = \sqrt{(d_{\text{od,GT}} - d_{\text{od,pred}})^2}.$$

Positional differences between the GT and predicted contact-point patterns are measured by the CD between the corresponding contact point sets M_{GT} and M_{pred}

$$\mathcal{L}_{\text{cp,pos}} = \mathcal{L}_{\text{cd}}(M_{\text{GT}}, M_{\text{pred}}).$$

Deviations in the number of contact points are captured by

$$\mathcal{L}_{\text{cp,num}} = \sqrt{(|M_{\text{GT}}| - |M_{\text{pred}}|)^2},$$

where $|M_{\text{GT}}|$ and $|M_{\text{pred}}|$ denote the number of contact points for the GT and the prediction, respectively.

Proximal contact metric: To capture proximal-contact behavior explicitly, the evaluation is extended by the proximal contact loss $\mathcal{L}_{\text{adj}}$ defined in (2.32). This metric quantifies the absolute penetration and vicinity situation between the target tooth and its adjacent teeth, thereby providing a direct measure of how precisely the reconstructed crown fits

into the interdental gap. It complements the $\mathcal{L}_{\text{cIoU}}$ metric, which reflects overall spatial overlap but does not specifically emphasize contact regions.

2.7 Evaluation

This section summarizes the evaluation protocol and reporting conventions used throughout the results. It specifies the training setup, the quantitative aggregation and visualization strategy, as well as the qualitative rendering settings to ensure consistent interpretation across models and tooth ranges. The model performance is evaluated in a two stage process. First, the reconstruction quality is assessed on a quantitative level using the metrics defined in Section 2.6 followed by a statistical analysis to assess significance. Second, qualitative results are presented to visually assess the reconstruction quality and functional properties of the generated teeth. The evaluation includes:

- Comparison of TST against common baseline PCD completion models.
- Model performance across different pose stages: initial prediction, after TPP, and after TPO.
- Model performance across different context sizes: full dentition, quadrant data, and adjacent data.
- Ablation studies analyzing the impact of different architectural and training choices.

2.7.1 Training

The networks are trained and evaluated on tooth ranges 1–3 and 4–6, respectively. These ranges correspond to different tooth types with distinct morphological characteristics and clinical requirements. Range 1–3 includes incisors and canines (anterior teeth), and range 4–6 corresponds to premolars and molars (posterior teeth).

For the non-sequential baseline models (PoinTr [30], PCN [47]), training is carried out on the same dataset using the objective defined in Section 2.3.6. In contrast to the sequence-based setting, all observed teeth are merged into a single partial input PCD by concatenating their points, yielding $|\mathcal{P}_{\text{partial}}| = \max\left(1024 \cdot |\mathcal{F}_{\text{obs}}|,\ 16384\right)$. Hyperparameters for all models are tuned on the validation set via grid search. Unless stated otherwise, all quantitative results are reported on the held-out test set using the metrics

defined in Section 2.6 and TST results are reported as the whole pipeline output after TPO refinement.

2.7.2 Representation

Results for model comparisons are reported as median values over all test samples with lower values indicating better performance for all evaluated metrics. Aligning with common conventions [30, 31, 49–51], units are omitted in the results for brevity. The units of the used metrics are provided in Table 2.2. The occlusion-specific metrics, as defined in Section 2.6.2, show different value ranges depending on the specific contact situation. Cases with contact points in the GT and no contact points in the prediction and vice versa are assigned a fixed penalty value of 100. Cases without any contact points in both the GT and the prediction are assigned a value of 0 for all occlusion-specific metrics. Due to the median representation of the results, the final reported values are not affected by the specific penalty values. The best performing model per metric is highlighted in bold. Columns without bold values indicate ties between multiple configurations (Tables 3.1, 3.2, 3.3, 3.4).

Each table is additionally accompanied by a box-plot visualizing the distribution of the metric values. For visualization, each metric is independently normalized to the range $[0, 1]$ across all groups shown in a given plot. Penalty values are normalized to 1.1 to enable statistical evaluation of all cases (Figures 3.1, 3.2, 3.3).

Statistically significant differences are indicated by asterisks (*). Significant results with statistical power below 0.8 are additionally highlighted in red. The specific pairwise comparisons follow the groupings indicated in the respective figures and captions.

2.7.3 Statistical evaluation

Analyses of the data distributions are conducted using Shapiro-Wilk tests ($p < 0.05$) to assess normality and Levene's test ($p < 0.05$) to evaluate homogeneity of variances across groups. Due to non-normality of the results, non-parametric statistical tests are employed. Groupwise statistical analysis is performed using the two-sided Wilcoxon signed-rank test with significance level $\alpha = 0.05$ to compare paired samples from different models. The null hypothesis states that there is no difference between the paired samples. Results with $p < \alpha$ are further evaluated with a one-sided test to determine the direction of

the effect. Statistical power is simulated with $n_{\text{sim}} = 1000$ bootstrap iterations. As the experiments are hypothesis-driven with a limited and predefined set of group comparisons, no correction for multiple testing is applied.

2.7.4 Qualitative results

Qualitative results are visualized using images of the reconstructed tooth meshes with color-coded contact distances to the antagonist teeth d_{anta}. The color map ranges from red (penetration) to blue (gap) with fixed thresholds of $-0.2\,\text{mm}$ and $0.2\,\text{mm}$ for all visualizations. Areas without occlusal contact are shown in white. PCD visualizations use a height dependent color map for better depth perception. For all visualizations, the same patient is used to enable comparability across all shown configurations (Figures 3.4, 3.5, 3.8).

2.7.5 Implementation details

This section lists the hyperparameters and dataset settings active in the pipeline. The values reflect the effective configuration used by each stage. Table 2.3 summarizes the hyperparameters for TST, TPP and TPO. All models are implemented in PyTorch and trained on 2×Nvidia RTX 4090 GPUs with 24 GB of VRAM.

To reduce the data processing overhead during training, RAM caching of preprocessed per-patient data is employed. Per-patient caches hold: (i) full set of segmented teeth with normals of shape $\mathbb{R}^{|\mathcal{F}_{\text{full}}|\times 2\times n_{\text{GT}}\times 3}$; (ii) cropped per tooth antagonists of shape $\mathbb{R}^{|\mathcal{F}_{\text{full}}|\times 2\times n_{\text{anta}}\times 3}$; (iii) per-tooth transforms of shape $\mathbb{R}^{|\mathcal{F}_{\text{full}}|\times 8}$. Missing and target teeth are filled with zeros. Observed teeth are downsampled to $|\mathcal{P}_{\text{obs},f}|$ points using FPS to reduce memory consumption. Target teeth PCDs are additionally stored as a single higher-resolution point set shape $\mathbb{R}^{n_{\text{GT}}\times 3}$ for objective evaluation during training. Indexing of teeth is performed via boolean masks for observed, adjacent, antagonist and target teeth. For TPO inference, RAM caching is disabled and higher-resolution point sets are loaded on demand. While the higher resolution data cannot be used for training due to GPU memory constraints, it provides more accurate surface representations of the antagonist and adjacent teeth that are crucial for a precise contact evaluation during optimization.

The number of predicted points $|\mathcal{P}_{\text{pred}}| = |\mathcal{P}_{\text{GT}}|$ is set to 2048 to balance reconstruction detail (Fig. 2.6) and computational efficiency.

Table 2.3: Implementation details for TST, TPP and TPO.

Component	Parameter	Value
	TST	
Model	Embedding width	$C = 768$
	Queries	$Q = 256$
	Progressive expansion steps	$n_{\text{pe}} = 3$
	Predicted points	$\lvert\mathcal{P}_{\text{pred}}\rvert = 2048$
Contact feature	Occlusion threshold	$\vartheta_{\text{occ}} = 0.2$
	Max. contact points	$\lvert\mathcal{P}_{\text{cp,fps}}\rvert = 128$
Transformer	Encoder / decoder blocks	6
	Attention heads	6
Loss	Curvature weight	$\beta_{\text{ccd}} = 0.5$
Optimizer	Type	AdamW [73]
	Learning rate	5×10^{-5}
	Weight decay	10^{-4}
Scheduler	Decay step	50 epochs
	Decay factor	0.7
	Minimum decay	0.02
Training	Batch size	6 (across 2 GPUs)
	Max. epochs	200
	Grad-norm clipping	10
Dataset	Per tooth observed points	$\lvert\mathcal{P}_{\text{obs},f}\rvert = 1024$
	Antagonist points	$\lvert\mathcal{P}_{\text{anta,crop}}\rvert = 1024$
	Antagonist crop threshold	$\vartheta_{\text{anta}} = 1\,\text{mm}$
	TPP	
Model	Embedding width	$C' = 128$
Optimizer	Type	AdamW [73]
	Learning rate	10^{-4}
	Weight decay	10^{-3}
Scheduler	Decay step	50 epochs
	Decay factor	0.7
	Minimum factor	0.02
Training	Batch size	256
	Max. epochs	200
	Grad-norm clipping	10
	TPO	
Variables	Scale clamp	$s_{\text{opt}} \in [0.8, 1.2]$
Surface reconstruction	Type	NKSR
	Detail level	0.95
	MISE iterations	1
	Grid upsample	1
Contact thresholds	Gate threshold	$\vartheta_{\text{gate}} = 0.1$
Loss	Weights	All 1.0
Optimizer	Type	SGD
	Learning rate	10^{-2}
	Max. iterations	$n_{\text{opt}} = 100$
Optimizer Dataset	Per tooth observed points for loss term evaluation	$\lvert\mathcal{P}_{\text{obs},f}\rvert = 4096$

3 Results

This chapter presents a comprehensive evaluation of the proposed TST method for domain guided functional tooth reconstruction. The network is evaluated quantitatively and qualitatively against two commonly employed PCD completion baselines. The impact of varying input context sizes and pose settings on reconstruction quality is investigated. In addition, an ablation study assesses the effectiveness of the proposed domain-specific model components.

3.1 Model and configuration comparisons

Table 3.1 reports the median test-set metrics for TST with TPO, PoinTr, and PCN on tooth ranges 1–3 and 4–6. Fig. 3.1 complements these medians by visualizing the full metric distributions and the corresponding pairwise statistical comparisons for the groups (TST/PoinTr, TST/PCN).

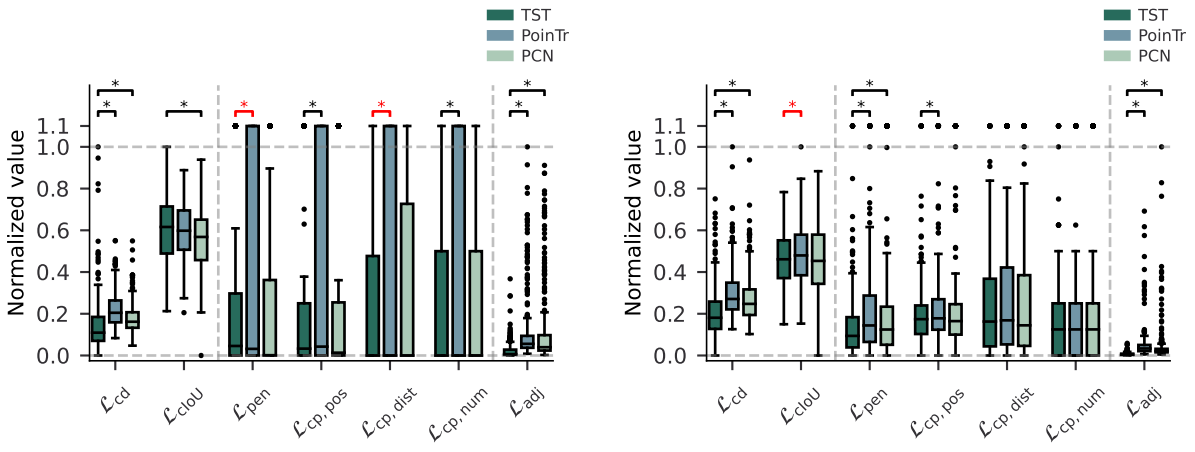
For range 1–3, TST attains statistically significant lower $\mathcal{L}_{\text{cd}}$ (0.250) and $\mathcal{L}_{\text{adj}}$ (0.064) values than both baselines. PCN achieves statistically significant lower $\mathcal{L}_{\text{cIoU}}$ (0.365) values than TST (0.396). Although the median $\mathcal{L}_{\text{cp,dist}}$ and $\mathcal{L}_{\text{cp,num}}$ are identical (0.000) across all models, Fig. 3.1 shows statistically significant differences between the TST and PoinTr distributions for all occlusion-specific metrics $\mathcal{L}_{\text{pen}}$, $\mathcal{L}_{\text{cp,pos}}$, $\mathcal{L}_{\text{cp,dist}}$, and $\mathcal{L}_{\text{cp,num}}$. The significances in $\mathcal{L}_{\text{pen}}$ and $\mathcal{L}_{\text{cp,dist}}$ exhibit low statistical power < 0.8. PCN yields the lowest median values for $\mathcal{L}_{\text{cp,pos}}$ (0.080) and $\mathcal{L}_{\text{pen}}$ (0.000), but these differences are not significant relative to TST ($\mathcal{L}_{\text{cp,pos}} = 0.197$, $\mathcal{L}_{\text{pen}} = 0.016$).

For range 4–6, TST again shows statistically significant lower values for $\mathcal{L}_{\text{cd}}$ (0.269), $\mathcal{L}_{\text{pen}}$ (0.041), and $\mathcal{L}_{\text{adj}}$ (0.068) than both baselines with overall higher values compared to range 1–3. The TST contact point position loss $\mathcal{L}_{\text{cp,pos}}$ (1.149) differs significantly from PoinTr (1.175). PCN shows significantly better $\mathcal{L}_{\text{cIoU}}$ (0.289) than TST (0.293), with low power < 0.8 and higher scattering. PCN also yields the lowest median values for $\mathcal{L}_{\text{cp,pos}}$ (1.087) and $\mathcal{L}_{\text{cp,dist}}$ (1.104), without significant differences relative to TST ($\mathcal{L}_{\text{cp,pos}} = 1.149$,

$\mathcal{L}_{\text{cp,dist}} = 1.242$). All models show similar performance in $\mathcal{L}_{\text{cp,dist}}$ and identical median performance for $\mathcal{L}_{\text{cp,num}}$ (1.000), without statistical differences.

Table 3.1: Median test-set metrics for TST with TPO, PoinTr, and PCN on ranges 1–3 and 4–6. Lower is better.

Configuration		Evaluation metrics (median)						
Model	Range	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	$\mathcal{L}_{\text{cp,num}}$	$\mathcal{L}_{\text{adj}}$
TST	1–3	**0.250**	0.396	0.016	0.197	0.000	0.000	**0.064**
PoinTr	1–3	0.356	0.384	0.011	0.255	0.000	0.000	0.479
PCN	1–3	0.308	**0.365**	**0.000**	**0.080**	0.000	0.000	0.344
TST	4–6	**0.269**	0.293	**0.041**	1.149	1.242	1.000	**0.068**
PoinTr	4–6	0.337	0.304	0.062	1.175	1.286	1.000	0.442
PCN	4–6	0.319	**0.289**	0.054	**1.087**	**1.104**	1.000	0.286



(a) Range 1–3.

(b) Range 4–6.

Fig. 3.1: Normalized box-plots corresponding to Table 3.1. Asterisks (*) indicate significant differences compared to TST (Wilcoxon, $\alpha = 0.05$, red: power < 0.8). Lower is better.

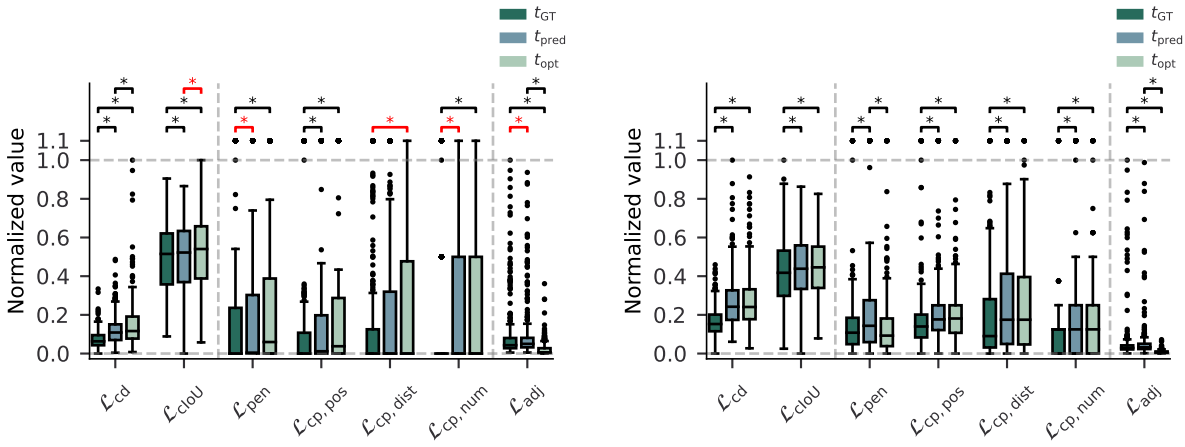
Table 3.2 reports the median test-set metrics for TST across the pose stages t_{GT}, t_{pred}, and t_{opt} on ranges 1–3 and 4–6. Fig. 3.2 complements these medians by showing the metric distributions and the corresponding pairwise statistical comparisons between all pose stages with comparison groups ($t_{\text{GT}}/t_{\text{pred}}$, $t_{\text{GT}}/t_{\text{opt}}$, $t_{\text{pred}}/t_{\text{opt}}$).

For range 1–3, the GT pose stage t_{GT} achieves significantly better morphological losses $\mathcal{L}_{\text{cd}}$ (0.190), $\mathcal{L}_{\text{cIoU}}$ (0.382) and occlusion-specific losses $\mathcal{L}_{\text{pen}}$ (0.000), $\mathcal{L}_{\text{cp,pos}}$ (0.000), and $\mathcal{L}_{\text{cp,num}}$ (0.000) than the remaining stages. The significances between t_{GT} and t_{pred} for $\mathcal{L}_{\text{pen}}$ and $\mathcal{L}_{\text{cp,num}}$ show low statistical power < 0.8. For $\mathcal{L}_{\text{cp,dist}}$, statistically significant

differences with low power (< 0.8) are observed between t_{GT} (0.000) and t_{opt} (0.000). These underpowered effects are consistent with the medians being identical across all pose stages for $\mathcal{L}_{\text{cp,dist}}$ (0.000) and $\mathcal{L}_{\text{cp,num}}$ (0.000). The optimized pose stage t_{opt} yields significantly lower $\mathcal{L}_{\text{adj}}$ values (0.064) than both prior stages t_{GT} (0.378) and t_{pred} (0.437), with a reduction factor of ≈ 6.

Table 3.2: Median test-set metrics for TST across pose stages t_{GT}, t_{pred}, and t_{opt} on ranges 1–3 and 4–6. Lower is better.

Configuration		Evaluation metrics (median)						
Pose	Range	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	$\mathcal{L}_{\text{cp,num}}$	$\mathcal{L}_{\text{adj}}$
t_{GT}	1–3	**0.190**	**0.382**	**0.000**	**0.000**	0.000	0.000	0.378
t_{pred}	1–3	0.241	0.386	0.001	0.062	0.000	0.000	0.437
t_{opt}	1–3	0.250	0.396	0.016	0.197	0.000	0.000	**0.064**
t_{GT}	4–6	**0.212**	**0.279**	0.047	**0.887**	**0.640**	1.000	0.311
t_{pred}	4–6	0.269	0.290	0.063	1.120	1.241	1.000	0.348
t_{opt}	4–6	0.269	0.293	**0.041**	1.149	1.242	1.000	**0.068**



(a) Range 1–3.

(b) Range 4–6.

Fig. 3.2: Normalized box-plots corresponding to Table 3.2. Asterisks (*) indicate significant differences between pose stages t_{GT}, t_{pred}, and t_{opt} (Wilcoxon, $\alpha = 0.05$, red: power < 0.8). Lower is better.

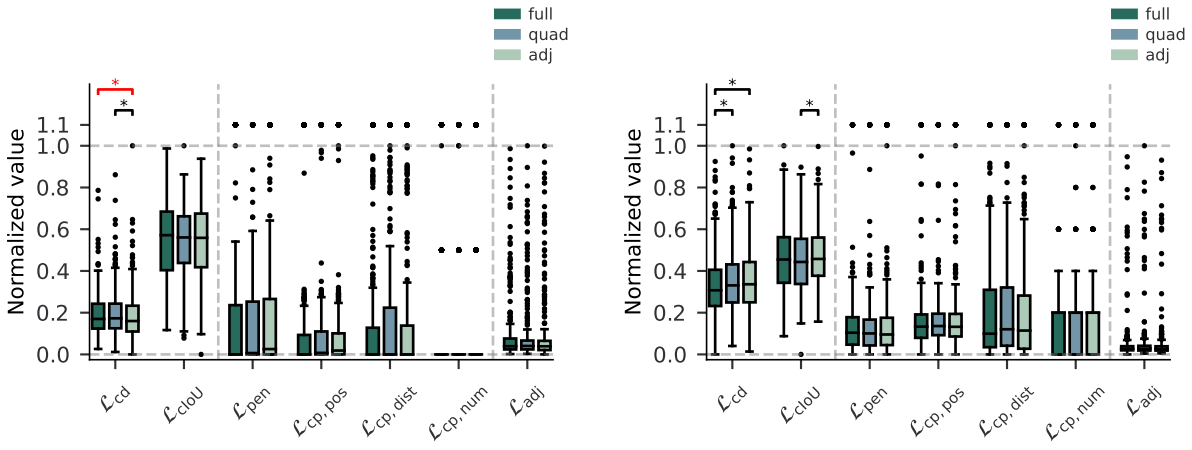
For range 4–6, t_{GT} again yields significantly lower morphological losses $\mathcal{L}_{\text{cd}}$ (0.212) and $\mathcal{L}_{\text{cIoU}}$ (0.279), as well as occlusion-specific metrics $\mathcal{L}_{\text{cp,pos}}$ (0.887), $\mathcal{L}_{\text{cp,dist}}$ (0.640), and $\mathcal{L}_{\text{cp,num}}$ (1.000), than both other stages. Although the median $\mathcal{L}_{\text{cp,num}}$ values are identical across all stages (1.000), the statistical analysis reveals high power (> 0.8) results. The optimized pose stage t_{opt} again achieves significantly lower $\mathcal{L}_{\text{adj}}$ values (0.068) than both

prior stages t_{GT} (0.311) and t_{pred} (0.348) with a reduction factor of ≈ 5, and it yields significantly lower $\mathcal{L}_{\mathrm{pen}}$ values (0.041) than t_{pred} (0.063).

Table 3.3 summarizes the median test-set metrics for TST trained with different context sizes (full, quad, adj) on ranges 1–3 and 4–6. Fig. 3.3 complements these medians by showing the metric distributions and the corresponding pairwise statistical comparisons between the context settings with comparison groups full–quad, full–adj, and quad–adj.

Table 3.3: Median test-set metrics for TST trained with different context sizes (full, quad, adj), evaluated on raw model outputs. Lower is better.

Configuration		Evaluation metrics (median)						
Context	Range	$\mathcal{L}_{\mathrm{cd}}$	$\mathcal{L}_{\mathrm{cIoU}}$	$\mathcal{L}_{\mathrm{pen}}$	$\mathcal{L}_{\mathrm{cp,pos}}$	$\mathcal{L}_{\mathrm{cp,dist}}$	$\mathcal{L}_{\mathrm{cp,num}}$	$\mathcal{L}_{\mathrm{adj}}$
full	1–3	0.190	0.382	**0.000**	**0.000**	0.000	0.000	0.378
quad	1–3	0.191	0.377	0.002	0.045	0.000	0.000	0.389
adj	1–3	**0.185**	**0.376**	0.007	0.111	0.000	0.000	**0.376**
full	4–6	**0.212**	0.279	0.047	0.887	**0.640**	1.000	**0.311**
quad	4–6	0.219	**0.273**	0.046	0.906	0.777	1.000	0.322
adj	4–6	0.221	0.281	**0.043**	**0.883**	0.737	1.000	0.317



(a) Range 1–3.

(b) Range 4–6.

Fig. 3.3: Normalized box-plots corresponding to Table 3.3. Asterisks (*) indicate significant pairwise differences between context settings (Wilcoxon, $\alpha = 0.05$, red: power < 0.8). Lower is better.

For range 1–3, all three context settings show similar performance across the evaluated metrics. Statistically significant improvements are observed only for $\mathcal{L}_{\mathrm{cd}}$ between adj (0.185) and the other context sizes full (0.190) and quad (0.191), with low statistical power < 0.8 for the comparison between adj and full.

For range 4–6, full achieves statistically significant better $\mathcal{L}_{\text{cd}}$ values (0.212) than the remaining context sizes quad (0.219) and adj (0.221). For $\mathcal{L}_{\text{cIoU}}$, quad (0.273) performs significantly better than adj (0.281) and attains the lowest overall median. The occlusion-specific metrics are not significantly affected by the context size.

3.2 Qualitative results

While the quantitative evaluation in Section 3.1 focuses on numerical metrics, this section provides qualitative reconstruction results to visually assess morphology and contact situations for an exemplary unseen patient, following the visualization protocol described in Section 2.7.4.

Fig. 3.4 and Fig. 3.5 show reconstructions for anterior (FDI 11–13, 31–33) and posterior (FDI 24–26, 44–46) teeth, respectively. The figures compare TST placed in the GT pose (TST_{GT}) and after TPP and TPO (TST_{TPO}) with the baselines PoinTr and PCN, and with the GT mesh.

For range 1–3, TST closely matches the GT morphology and contact situation. Consistent with Table 3.2 and Fig. 3.2, TPO does not reproduce the contact accuracy achieved in the GT pose. In terms of morphology, TST provides stable reconstructions across the anterior teeth. For teeth 1–2, PoinTr and PCN produce acceptable shapes. Starting at tooth 3 and with increasing anatomical complexity toward the posterior region, both baselines show oversmoothed surfaces and reduced fine-detail fidelity compared to TST and the GT.

This behavior becomes more pronounced in Fig. 3.5. Both baselines exhibit clearer morphological artifacts, including irregular surfaces, unnatural spikes, and holes. The severity of these artifacts increases with complexity from tooth 4 to 6. While the contact situation remains acceptable for teeth 4–5 in PoinTr and PCN, tooth 6 shows substantial penetration into the antagonists. This observation is consistent with Table 3.1 and Fig. 3.1, where TST outperforms both baselines in $\mathcal{L}_{\text{pen}}$ for range 4–6. Across the posterior teeth, TST yields visually plausible morphologies and contact situations, including improved contact point distributions and reduced penetration relative to both baselines. The pose-stage comparison reproduces the trends from Table 3.2 and Fig. 3.2. The GT pose shows the best contact point distribution, while TPO achieves the best penetration behavior.

Regarding morphological details, TST preserves fine-scale structure across the posterior teeth. With increasing tooth complexity, smoothing effects become more apparent, particularly in occlusal regions of tooth 6 where fissures appear less distinct than in the GT.

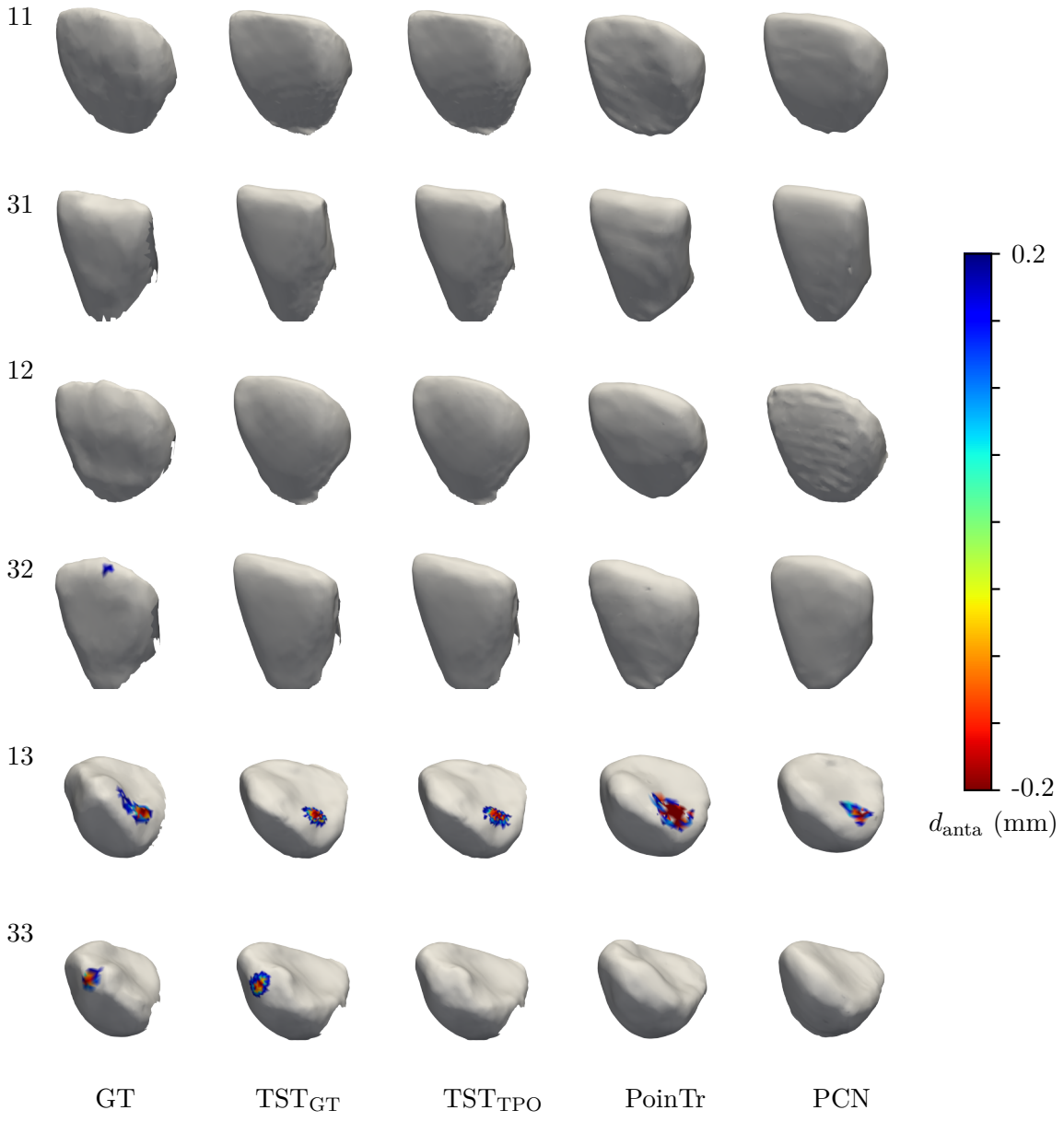


Fig. 3.4: Anterior reconstructions (FDI 11–13, 31–33) comparing TST and baselines; colors encode the distance to the antagonist teeth d_{anta} (red: penetration, blue: gap). TST_{GT}: TST prediction with GT-pose; TST_{TPO}: with TPP and TPO; PoinTr and PCN: baseline networks trained on the same data.

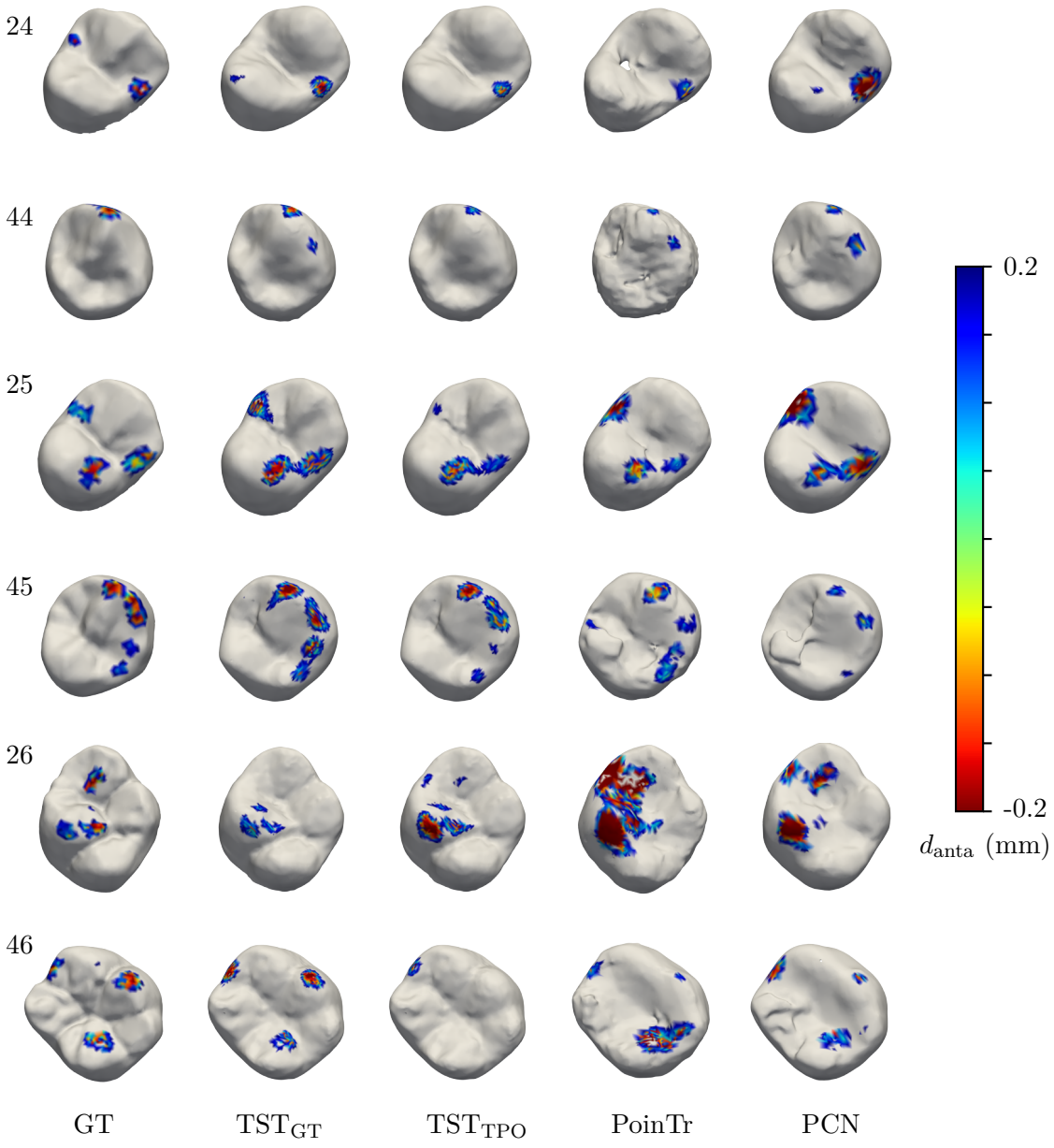


Fig. 3.5: Posterior reconstructions (FDI 24–26, 44–46) comparing TST and baselines; colors encode the distance to the antagonist teeth d_{anta} (red: penetration, blue: gap). TST_{GT}: TST prediction with GT-pose; TST_{TPO}: with TPP and TPO; PoinTr and PCN: baseline networks trained on the same data.

Fig. 3.6 shows the reconstruction results for TST with TPO in the adjacent teeth context for both anterior and posterior teeth. The figure validates the low $\mathcal{L}_{\text{adj}}$ values, reported in Table 3.2 and Fig. 3.2, by demonstrating a visually plausible fit of the predicted tooth shapes into the gap between the neighboring teeth.

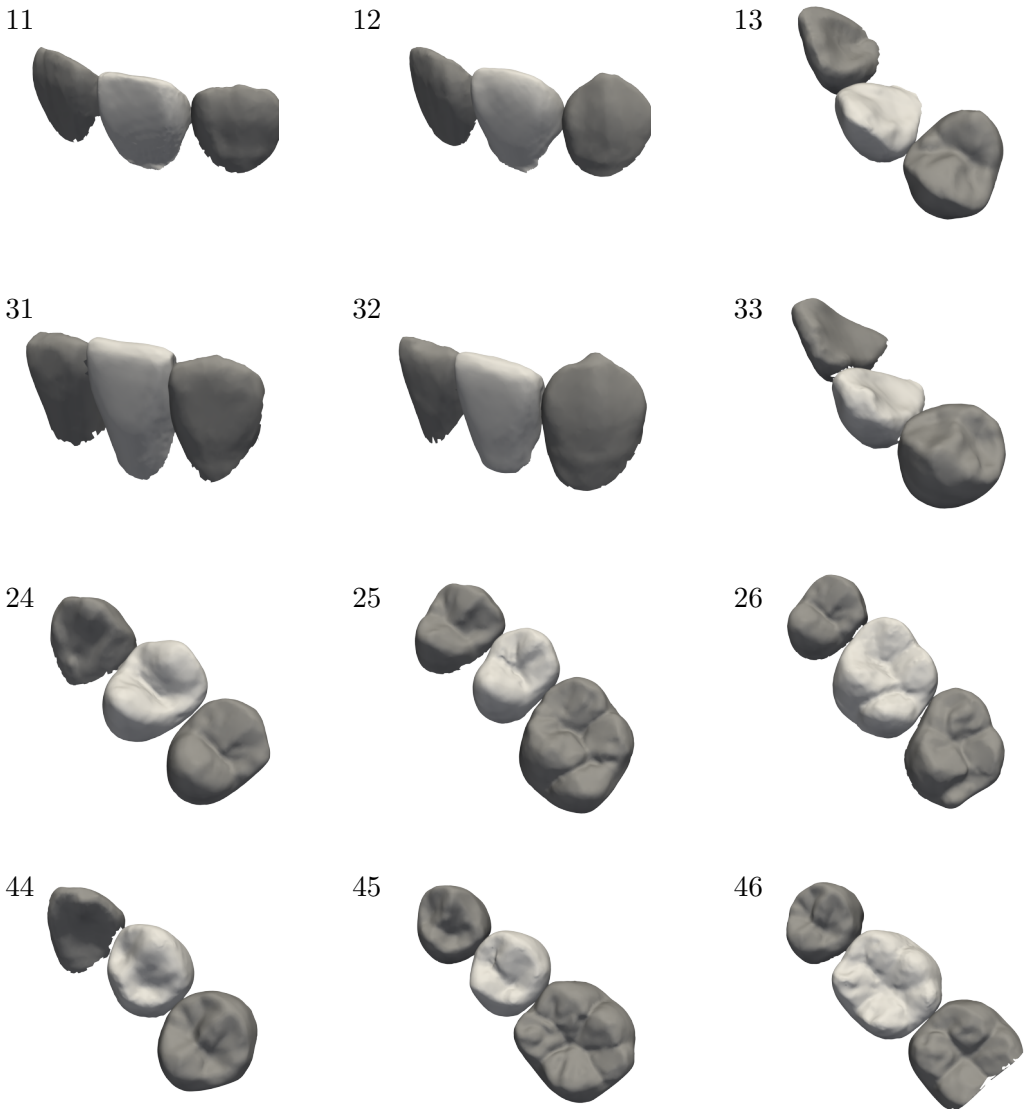


Fig. 3.6: Reconstructions for TST_{TPO} (Fig. 3.4, Fig. 3.5) in the adjacent teeth context. Target teeth are shown in white, while neighboring teeth are depicted in gray.

To visually contextualize the reconstruction-quality differences between TST and the baselines, Fig. 3.7 shows the raw PCD outputs of all models for teeth 1–6. TST produces a dense and largely uniform point distribution that closely follows the tooth surface. In contrast, both baselines exhibit characteristic artifacts. PoinTr forms scattered point clusters near the centroid of the target tooth, which yields an uneven sampling density and reduces the effective spatial resolution over the remaining surface. In addition, PoinTr and PCN generate noisier point sets with points scattered around the surface, which promotes surface-reconstruction artifacts and contributes to the loss of morphological detail

observed in Fig. 3.4 and Fig. 3.5.

The mesh-reconstruction setup used for these figures is already adapted to the noise characteristics of the baseline outputs by selecting a lower NKSR detail level than for TST, as described in Section 2.5.1. Fig. 2.7 illustrates the impact of these artifacts when the reconstruction parameters are not adjusted.

Table 3.2 and Fig. 3.2 indicate that the training context size has only a minor influence on the quantitative reconstruction results. Fig. 3.8 provides a qualitative comparison for TST trained with different context sizes (full, quad, adj) on an exemplary unseen patient. The figure includes anterior (FDI 11–13, 31–33) and posterior (FDI 24–26, 44–46) teeth and visualizes contact points as described in Section 2.7.4.

Overall, all three context sizes yield visually similar morphologies and contact situations. With increasing tooth complexity, an observable trend toward improved morphological detail and smoother surfaces emerges as the available context increases. This is most apparent when comparing full and adjacent-only context for the molars (FDI 26, 46). The contact situations remain visually comparable across all context sizes. These observations are consistent with Table 3.3 and Fig. 3.3, where the full context achieves significantly better $\mathcal{L}_{\text{cd}}$ values than the remaining context sizes for range 4–6, while the occlusion-specific metrics show no statistically significant differences.

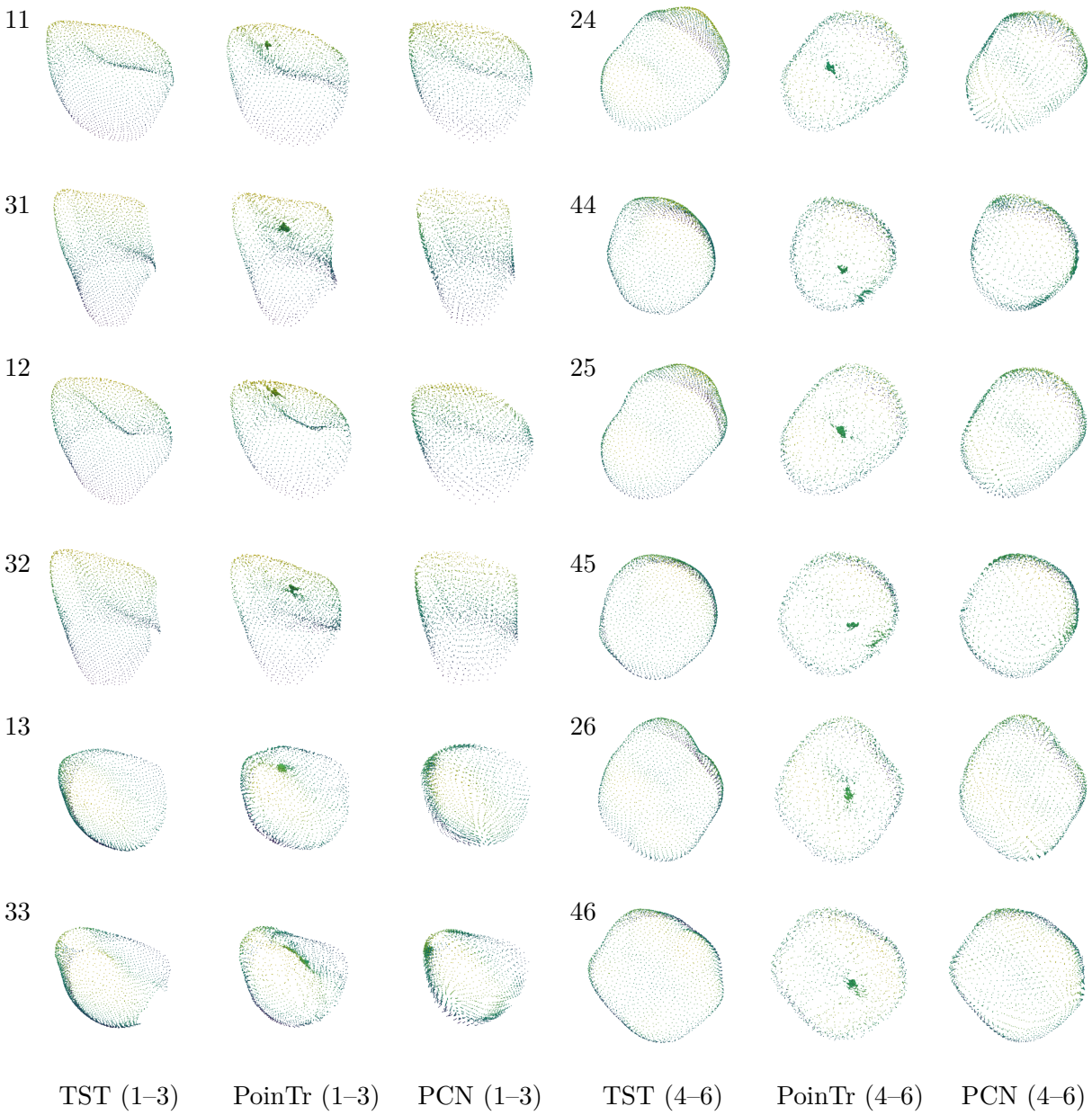


Fig. 3.7: Raw model prediction output of the mesh reconstructions from Fig. 3.4 and Fig. 3.5. The respective tooth ranges are indicated after the model names (1–3: anterior teeth, 4–6: posterior teeth). As the tooth morphology is not affected by the TPP, the TST PCDs are visualized before pose prediction ($\mathrm{TST}_{\mathrm{GT}}$).

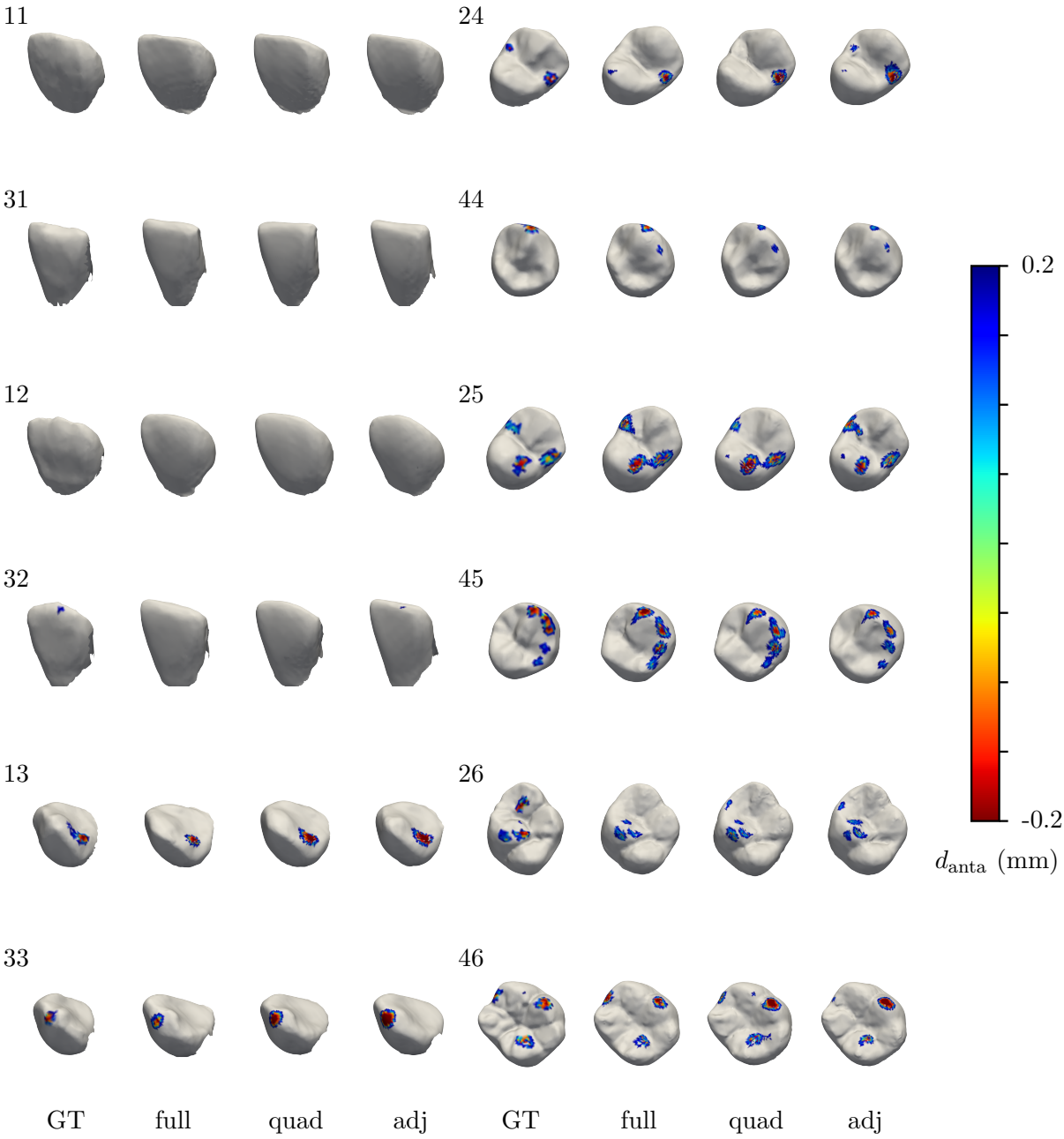


Fig. 3.8: Reconstructions for TST trained with different context sizes (full, quad, adj); colors encode the distance to the antagonist teeth d_{anta} (red: penetration, blue: gap). Left: anterior teeth (FDI 11–13, 31–33); right: posterior teeth (FDI 24–26, 44–46).

3.3 Ablation studies

To assess the impact of key design decisions in the architecture and training setup, an ablation study is performed on the principal variable components (Table 3.4). The baseline configuration A corresponds to the full TST model described in Chapter 2. Each ablation modifies exactly one aspect of this baseline, and the corresponding model is retrained under identical conditions for 100 epochs. Evaluation is conducted on the validation set, reporting median metrics from the best-performing checkpoint selected by the $\mathcal{L}_{\text{cd}}$ validation loss. To compare ablations, the mean rank across all reported metrics is computed, where lower values indicate better overall performance. Configurations that match or outperform the baseline in mean rank are highlighted. For qualitative reference, posterior-tooth reconstructions are shown for all ablations in Fig. 3.9.

Table 3.4a summarizes architectural variants, ranging from alternative encoders to the removal of domain-specific components. Replacing the PCD encoder in the morphological and anatomical tooth encoder from DGCNN to a commonly used PointNet [30, 45] style architecture improves the overall quantitative metrics. The qualitative comparison, however, indicates increased surface-adjacent noise for the PointNet variant, which can propagate into artifacts in the reconstructed mesh. Removing the CPM, substituting folding-based upsampling [48], or disabling the anatomical priors consistently reduces performance. The anatomical priors comprise the FDI embedding as well as adjacent and antagonist embeddings.

Table 3.4b reports ablations of the data normalization pipeline affecting orientation and scale. Disabling adjacent rotation normalization yields quantitative results similar to the baseline, but produces noisier surface predictions, as also visible in Fig. 3.9. To evaluate the proposed reversible arch-based scale normalization, the baseline is replaced by tooth-level scaling that normalizes each tooth using the scale of all teeth with identical FDI codes across the dataset. In addition, scaling is disabled entirely. Both alternatives reduce performance, suggesting that arch-based normalization effectively reduces inter-patient variability while preserving relative size relationships within the dental arch.

To evaluate the curvature-weighted chamfer loss, Table 3.4c compares curvature weights with $\beta_{\text{ccd}} \in \{0.0, 0.5, 1.0\}$. The setting $\beta_{\text{ccd}} = 0$ corresponds to the standard chamfer loss without curvature weighting. The quantitative results indicate no improvement from curvature weighting. The qualitative examples in Fig. 3.9 suggest that mild weighting

($\beta_{\text{ccd}} = 0.5$) yields slightly more consistent surface representations. The observed effect remains small relative to the architectural and data-related ablations.

Table 3.4: Ablation studies for TST (data range: adjacent + antagonist, tooth range: 4–6) trained for 100 epochs. Lower is better for all metrics. Median validation metrics are reported. The final column in each subtable shows the mean rank across all listed metrics (lower is better). The baseline model A is identical across all subtables. The ablations cover (a) architectural variants, (b) data-related configuration choices, and (c) training loss variations.

(a) Model ablation. Encoder: implemented PCD encoder architecture (Section 2.3.1); CPM: contact point module (Section 2.3.2); Upsampling: point set upsampling method (Section 2.3.5); A. Prior: Use of anatomical priors (Section 2.3.1).

Configuration					Evaluation metrics (median)					
Model	Encoder	CPM	Upsampling	A. Prior	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	Rank
A	DGCNN [58]	✓	EXPANSION	✓	0.217	0.291	0.044	0.864	0.592	2.2
B	POINTNET [45]	-	-	-	0.216	0.276	0.039	0.846	0.683	**1.8**
C	-	×	-	-	0.220	0.294	0.045	0.808	0.597	2.4
D	-	-	FOLDING [48]	-	0.222	0.294	0.039	0.926	0.719	3.6
E	-	-	-	×	0.229	0.313	0.053	0.828	0.707	3.8

(b) Data transformation ablation (Section 2.2.3). Adj-Rot: adjacent rotation normalization; Scale: scale normalization; Scale-Type: type of scale normalization (arch: per patient scale across all teeth in arch; tooth: per tooth scale across all patients).

Configuration				Evaluation metrics (median)					
Model	Adj-Rot	Scale	Scale-Type	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	Rank
A	✓	✓	ARCH	0.217	0.291	0.044	0.864	0.592	**2.0**
F	×	-	-	0.223	0.281	0.039	0.821	0.662	**2.0**
G	✓	×	-	0.232	0.271	0.045	0.864	0.614	2.6
H	✓	✓	TOOTH	0.222	0.291	0.048	0.855	0.747	3.0

(c) Curvature chamfer weight β_{ccd} ablation (Section 2.3.6). A weight of $\beta_{\text{ccd}} = 0$ corresponds to the standard chamfer loss without curvature weighting.

Configuration	Evaluation metrics (median)					
Model	β_{ccd}	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	Rank
A	0.5	0.217	0.291	0.044	0.864	0.592	2.0
I	0.0	0.214	0.290	0.041	0.832	0.606	**1.2**
J	1.0	0.217	0.297	0.041	0.859	0.680	2.2

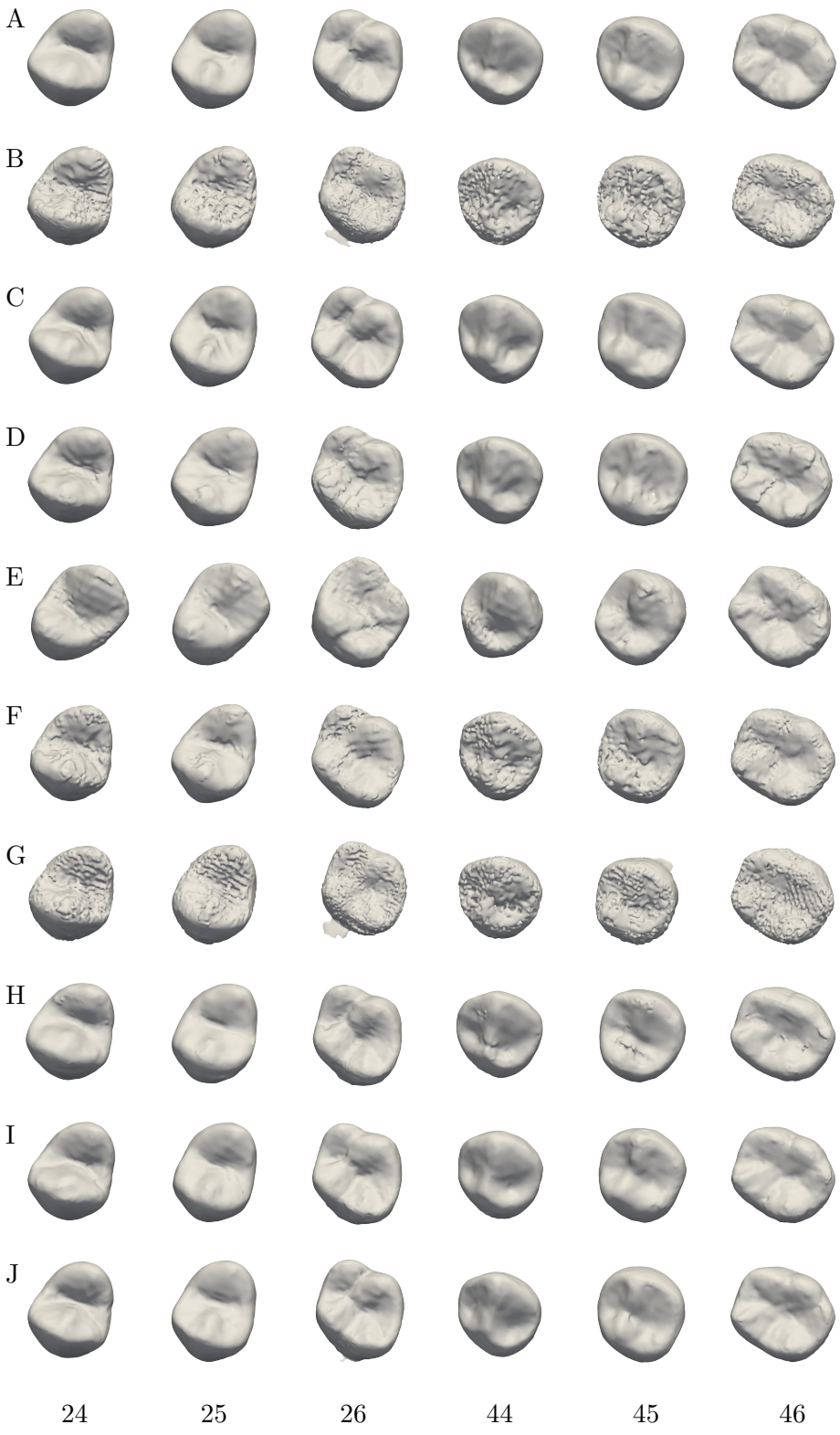


Fig. 3.9: Visual examples of all ablated model configurations (Table 3.4). The columns correspond to different teeth and the rows show all ablation configurations.

3.4 Performance

On a single NVIDIA RTX 4090 GPU, the end-to-end pipeline runtime can be decomposed into data loading and preprocessing (1 s), TST checkpoint loading and model initialization (3 s), TST and TPP inference including surface reconstruction (2 s), and TPO optimization (3 s). The resulting average runtime per reconstructed tooth, including surface reconstruction and TPO, is approximately 6 s. Model loading and initialization are performed once per session and are therefore excluded from the per-tooth inference time. The total training time for TST on tooth range 4–6 with full data context on two NVIDIA RTX 4090 GPUs is approximately 63 hours for 200 epochs.

4 Discussion

This chapter interprets the results in Chapter 3 in the context of the functional reconstruction pipeline described in Chapter 2. All interpretations follow the metric definitions in Section 2.6 and the reporting conventions in Section 2.7. Consistent with Section 2.7, TST denotes the end-to-end pipeline output after TPP and TPO unless stated otherwise.

Scope and contributions. This work formulates single-tooth crown reconstruction as a sequence-to-shape problem on segmented 3D tooth PCDs. The proposed TST architecture explicitly models the dentition as a structured tooth sequence with anatomical priors derived from FDI identity embeddings and role embeddings for adjacent, antagonist, missing, and target teeth (Section 2.3.1). Functional signals are incorporated at the per-tooth level via the CPM (Section 2.3.2), and the dense target PCD is generated in a coarse-to-fine manner through a transformer decoder and the progressive expansion module (Section 2.3.5). Beyond shape prediction, the pipeline includes a dedicated translation predictor (TPP) and a contact-aware refinement stage (TPO) that optimizes translation and scale under occlusal and proximal objectives (2.33). The evaluation in Chapter 3 therefore assesses not only geometric similarity, but also clinically motivated contact behavior and the practical implications of separating learned shape generation from controllable placement refinement.

Main quantitative findings. Across both tooth ranges, TST achieves statistically significant improvements in morphological fidelity and proximal fit relative to the completion baselines PoinTr and PCN. Specifically, TST yields lower $\mathcal{L}_{\mathrm{cd}}$ and substantially lower proximal contact loss $\mathcal{L}_{\mathrm{adj}}$ for both anterior teeth (1–3) and posterior teeth (4–6) (Table 3.1, Fig. 3.1), under the statistical protocol in Section 2.7.3. For posterior teeth, TST also improves antagonist penetration behavior, with significantly lower $\mathcal{L}_{\mathrm{pen}}$ than both baselines (Table 3.1, Fig. 3.1). In contrast, PCN attains lower $\mathcal{L}_{\mathrm{cIoU}}$ values (Table 3.1), which is consistent with $\mathcal{L}_{\mathrm{cIoU}}$ being computed on bounding boxes (Section 2.6) and therefore reflecting coarse spatial overlap rather than fine-scale anatomical agreement. As expected,

all models achieve better quantitative performance on anterior teeth than on posterior teeth due to the lower anatomical complexity of incisors and canines. Across both tooth ranges, TST exhibits similar relative performance trends with respect to morphology.

Several occlusion metrics exhibit identical or near-identical medians across groups (e.g., $\mathcal{L}_{\text{cp,dist}}$ and $\mathcal{L}_{\text{cp,num}}$ in Table 3.1), yet the corresponding box-plots reveal group-wise distribution shifts and pairwise differences for selected comparisons (Fig. 3.1). Since occlusion metrics include fixed penalties for contact failure edge cases, and medians can mask changes that occur primarily in the tails or in subsets of samples, distribution level analyses are essential to fully capture clinically relevant contact behavior. There is, however, no clear trend favoring one model across all occlusion metrics. This reflects the presence of strong functional priors that are implicitly encoded in the partial input data, which enables all models to achieve reasonable occlusal contact patterns even without dedicated functional feature extraction. This is consistent with the ablation results in Table 3.4a, where removing the CPM has only a minor effect on occlusion metrics while still improving morphological fidelity.

These findings validate the core design principles of TST. The sequence-to-shape formulation with anatomical priors effectively captures tooth-specific morphology across the dental arch and significantly improves detail recovery compared to generic shape-completion baselines. The quantitative results indicate that the morphological detail improvements are consistent across tooth types.

Qualitative fidelity beyond summary statistics. To validate the quantitative findings, contemporary mesh reconstructions are visualized in Fig. 3.4 and Fig. 3.5 for anterior and posterior teeth, respectively. The mesh visualizations illustrate that quantitative summary metrics do not fully capture clinically relevant fine-scale anatomy and surface artifacts. For simpler anterior teeth, all approaches can produce plausible crowns, whereas the posterior cases reveal pronounced differences. PoinTr and PCN increasingly oversmooth occlusal morphology and introduce artifacts such as holes, spikes, and irregular fissure patterns as anatomical complexity increases. Consistent with the quantitative results, the contact points visualized in Fig. 3.4 and Fig. 3.5 illustrate that all models are able to recover clinically plausible occlusal contact patterns from the partial antagonist input. In the most complex molar examples, baseline reconstructions can exhibit clinically unacceptable antagonist penetration (e.g., tooth 46 in Fig. 3.5), consistent with the posterior penetration differences in $\mathcal{L}_{\text{pen}}$ (Table 3.1). In contrast, TST more consistently preserves

high-frequency occlusal structure and yields visually plausible contact situations across the full tooth range. Posterior fissure depth and the finest surface detail can, however, appear attenuated relative to the GT meshes.

Contrary to the observed consistent morphological quantitative improvements, the visual quality differences are more pronounced for posterior teeth, where anatomical complexity amplifies the impact of surface artifacts and detail loss on clinical plausibility. This highlights the limitations of global summary metrics in capturing domain-specific fidelity aspects that directly translate into clinical usability of the reconstruction.

Failure modes from point-set prediction and surface reconstruction. The raw PCDs in Fig. 3.7 provide a mechanism-level perspective on these qualitative differences. TST produces a dense and largely uniform point distribution that follows the tooth surface, whereas PoinTr and PCN exhibit characteristic clustering and outliers. Because the CD training objective in Section 2.3.6 does not enforce uniform sampling, it can permit degenerate nearest-neighbor matchings that reduce distance values while yielding clustered or uneven point distributions. This is a known failure mode of chamfer-based PCD objectives [74]. These point-level artifacts interact with the downstream surface reconstruction stage. As demonstrated in Fig. 2.7, higher NKSR detail levels amplify noise-induced artifacts for PoinTr and PCN, which necessitates more conservative reconstruction settings and reduces recoverable high-frequency detail. This interaction explains why mesh quality can degrade even when differences in $\mathcal{L}_{\text{cd}}$ and especially $\mathcal{L}_{\text{cIoU}}$ appear moderate.

Posterior morphology detail limitations. The remaining gap of the TST reconstructions to GT anatomy for complex molars likely reflects both representation limits and surface reconstruction sensitivity. While Fig. 2.6 shows that NKSR can recover fine detail from well-behaved GT samples with the same number of sampled points, predicted PCDs inevitably contain noise and occasional outliers. Although TST produces relatively clean point sets, these artifacts are not explicitly addressed by the chamfer-based training objective and can be amplified by the surface reconstruction.

Plausible directions include adaptations to surface reconstruction and architectural refinements. Surface reconstruction could be tailored to better handle the noise characteristics of predicted PCDs, for example by training NKSR on domain-specific prediction data. Architectural refinements could further integrate surface prediction into the learning process through output representations that jointly predict surfaces and normals [34–37].

Pose stage ablation. The stage-wise structure of the pipeline is motivated by the clinical distinction between generating a morphologically plausible proposal and achieving a functionally feasible placement under contact constraints. While, in theory, one shot prediction with deterministic and unique shape and pose output is desirable, the highly non-linear and patient-specific nature of functional contacts renders this approach challenging in practice. Thus, the TPO objective in (2.33) explicitly encodes proximal and occlusal targets as well as minimal-thickness constraints to the preparation surface when available. To enable a flexible user interface, the weights and targets can be adjusted to support tighter or looser contact preferences depending on the use case. As with most non-convex objectives, initialization plays a key role in optimization success. TPP therefore serves as a dedicated pre-alignment stage to provide a favorable starting point for optimization.

The pose ablation in Table 3.2 clarifies the resulting trade-off. Using the GT pose t_{GT} isolates the shape predictor and yields the strongest morphology and occlusion contact-pattern metrics (Fig. 3.2), but it assumes access to a clinically unknown transformation. The optimized pose t_{opt} instead targets functional feasibility. Across both ranges, TPO yields substantially lower $\mathcal{L}_{\mathrm{adj}}$ than t_{GT} and t_{pred}, reducing the proximal-contact loss by approximately a factor of 5–6 (Table 3.2) and producing a visually plausible interdental fit (Fig. 3.6). For posterior teeth, it also improves penetration behavior relative to t_{pred} (Table 3.2).

Since adjacent contacts are not explicitly enforced during pose-agnostic shape prediction, the proximal contact situation of the initial shape prediction does not align with clinical requirements that do not allow penetrations or excessive gaps. The occlusal contact patterns are, however, already adapted to the antagonist input, as demonstrated by the occlusion metrics for t_{GT}. Consequently, the optimizer may need to deviate from poses that best preserve occlusal contact-point patterns in order to resolve proximal penetrations or excessive gaps. This explains why some occlusion-related metrics can degrade even as proximal alignment and overall penetration behavior improve (Table 3.2, Fig. 3.2). Reducing this trade-off likely requires incorporating adjacent contact cues into TST training without leaking absolute pose information and thus preserving pose-agnostic shape fidelity and generalization.

Context size ablation. The context-size comparisons in Table 3.3 and Fig. 3.3 indicate that reducing scan context from full-arch to quadrant or adjacent-only has only a mi-

nor influence on the reported occlusion metrics, while posterior morphology benefits from larger context. For range 4–6, full context yields significantly lower $\mathcal{L}_{\text{cd}}$ than quadrant and adjacent-only settings (Fig. 3.3b). These findings align with previous research on context effects in AI-based commercial crown design software [64]. Qualitatively, Fig. 3.8 suggests smoother and more detailed molar morphology with broader context, while antagonist-distance patterns remain visually comparable across contexts. This supports the interpretation that antagonist inclusion already captures the most relevant functional information reflected by the employed occlusion metrics, whereas broader same-arch context primarily contributes to morphological detail for high-complexity teeth. The increased morphological fidelity can be partially explained by the presence of the mirror tooth in the full context, which provides a strong morphological-anatomical prior for the target tooth shape. This interpretation is consistent with conventional dental design practice, where mirror-tooth information is frequently used to guide crown morphology [15]. This interpretation is bounded by the present metrics and does not exclude clinically relevant context effects that are not captured by the evaluation.

Model design ablations. The model design ablation study (Table 3.4, Fig. 3.9) reinforces that several domain-specific design choices contribute to stable reconstruction behavior and surface quality. Disabling anatomical priors (Section 2.3.1) worsens overall performance, supporting structured identity conditioning as an effective inductive bias beyond generic point-set completion. Removing the CPM does not substantially change the reported occlusion metrics in isolation, suggesting that multi-tooth geometry already carries strong functional priors. Nevertheless, the full configuration achieves the best aggregate behavior, and CPM remains a lightweight mechanism to inject explicit contact-related signals. The upsampling ablation further indicates that the staged, offset-based expansion in the PEM (Section 2.3.5) produces more regular PCDs than FoldingNet-style [48] deformation, aligning with the point-level artifacts observed for the baselines and previous research on point completion [51, 52, 66]. Data normalization choices also influence stability, with adjacent rotation normalization reducing surface noise and improving qualitative detail levels. This supports the interpretation that reducing inter-patient variability through anatomically meaningful normalization facilitates learning of fine-scale morphology. The proposed arch-based scale normalization further improves performance relative to tooth-level or no scaling. This indicates that preserving relative size relationships within the dental arch for each patient better supports anatomically plausible size predictions than normalizing each tooth independently across the dataset over all

patients.

Finally, the curvature-weighted chamfer loss was introduced to better preserve fine anatomical structures (Section 2.3.6), but the loss ablation shows no quantitative improvement and only minor qualitative effects (Table 3.4c). This suggests that, in the present setting, architectural and representation level factors dominate the preservation of high frequency detail. The previously discussed attenuated posterior fissure depth and the qualitative ablation results in Fig. 3.9 (A, I, J) indicate that the representation of fine-scale anatomical detail remains an architectural rather than purely a loss design challenge. A disciplined future direction is therefore to incorporate refinement mechanisms that explicitly target clinically relevant structures, for example through staged refinement focused on anatomical regions or landmark guided detail enhancement, while accounting for the limited availability of dental landmark annotations [62].

Runtime and workflow implications. The runtime analysis in Section 3.4 indicates practical feasibility of the full pipeline, with an average per-tooth runtime of approximately 6 s on a single NVIDIA RTX 4090 GPU and a loaded model instance. Because TPO accounts for roughly half of the per-tooth time, a practical workflow is to first generate an initial morphological proposal and to defer TPO until the user accepts the proposal or requests contact adjustments. This staged usage is consistent with the separation between learned proposal generation and controllable, constraint-based placement refinement encoded in (2.33).

Limitations and future work. From a statistical perspective, some pairwise comparisons exhibit low power (Fig. 3.1, Fig. 3.2, Fig. 3.3). Consequently, non-significant results should not be interpreted as evidence of equivalence, and significant differences should be interpreted together with reported power levels. From a construct-validity perspective, the employed quantitative metrics capture complementary aspects of morphology and contacts, but they do not fully quantify fine-scale anatomical fidelity and surface artifacts as demonstrated in the qualitative evaluation. Visual inspection of both meshes and raw PCDs therefore remains necessary (Section 2.7.4), but it is inherently subjective and limited in coverage.

From an internal- and external-validity perspective, the evaluation assumes accurate tooth segmentations and the availability of correctly oriented antagonist scans. It does not quantify sensitivity to segmentation errors or incorrect antagonist alignment. Moreover,

broader validation on heterogeneous clinical data from multiple scanners and patient populations remains necessary. A central scope boundary is that the dataset does not include preparation scans (Section 2.2). Therefore, the minimal-thickness constraints supported by TPO cannot be directly evaluated in the present experiments, and the pipeline should be interpreted primarily as producing an anatomically plausible outer crown proposal together with contact-aware placement refinement. Preparation geometry remains essential for final manufactured crown design. Thus, a practical path is to combine the proposed outer-shape reconstruction with conventional design tools for the inner crown surface, consistent with hybrid workflows discussed in prior work [12].

Future work should therefore

- incorporate preparation scans and evaluation protocols that directly assess thickness constraints and preparation fit,
- validate robustness across more diverse clinical data and scanning conditions,
- develop training strategies that improve initial proximal contacts without leaking pose information to reduce the required optimization corrections,
- improve preservation of high-frequency anatomy through representation-aware refinement and/or surface reconstruction methods better matched to predicted point set characteristics.

5 Summary

Objectives: Fully automated reconstruction of patient-specific tooth crowns from partial digital impressions remains challenging, as restorations must satisfy both anatomical shape requirements and functional occlusal and proximal contacts. This work addresses this gap by developing and evaluating a deep-learning-based tooth sequence pipeline that generates crown proposals from multi-tooth 3D scan context and refines their placement under explicit contact constraints.

Methods: The proposed TST treats crown reconstruction as a sequence-to-shape problem on segmented 3D tooth PCDs. Teeth are represented in a normalized adjacent teeth reference frame and augmented with tooth-identity information following the FDI numbering scheme. Each observed tooth PCD is encoded into a morphological feature embedding and fused with learned anatomical priors derived from its FDI code. Complementary role embeddings distinguish adjacent, antagonist, missing, and target teeth, enabling a unified sequence representation under varying context availability. To integrate patient-specific occlusal information at the per-tooth level, the CPM derives occluded surface points for each observed tooth with respect to the antagonist set, encodes a fixed number of sampled contact points, and injects the resulting contact features into the tooth sequence via cross-attention. The enriched sequence is then processed by a transformer encoder stack that alternates self-attention with nonlinear feature refinement through an MLP. From the final attention-refined target feature, the CPG predicts a sparse coarse point set and derives geometry-aware query features. A transformer decoder subsequently refines these queries via self-attention and cross-attention to the encoded tooth-context features. Finally, the offset-based PEM upsamples the coarse prediction in multiple stages. At each stage, per-point features are refined and an MLP predicts local offset vectors that spawn new points, yielding a dense target PCD with progressively added geometric detail. Training employs a curvature-weighted chamfer loss to emphasize high-curvature anatomical structures.

Because the patient-frame translation component of the partially reversible normalization transform cannot be inferred from tooth geometry alone, TPP predicts the target translation from the observed sequence of per-tooth translations. The translation input is stabilized by subtracting the mean translation of the target's adjacent teeth, embedded using a residual MLP, and combined with the same FDI-based identity and role embeddings used for TST. The pose predictor omits the full encoder stack and instead applies a cross-attention decoder that uses the target embedding as query to aggregate contextual information. A subsequent MLP head regresses the normalized translation, which is then transformed back to the patient frame. Training uses a mean-squared error objective.

TPO refines placement by optimizing an absolute translation, initialized from TPP, and an isotropic scale. The predicted PCD is converted to a watertight surface using NKSR after kNN-based normal estimation, and the resulting dense mesh vertices are used for contact evaluation against adjacent teeth, antagonists, and optionally a preparation surface. Penetration and near-contacts are quantified via orientation-aware nearest-neighbor distances with sequential mutual gating. The final objective enforces symmetric proximal contacts, matches prescribed proximal and occlusal target profiles, and maintains a minimal thickness margin to the preparation when available.

Using a combined set of morphology and contact metrics with distribution-aware statistical analysis, the end-to-end TST pipeline is evaluated against the common completion baselines PoinTr and PCN on tooth ranges 1–3 (anterior) and 4–6 (posterior).

Results: Across quantitative and qualitative experiments, TST demonstrates robust reconstruction quality compared to two commonly used PCD completion baselines (PoinTr and PCN). In particular, TST yields statistically significant improvements in morphological fidelity and proximal alignment, including lower $\mathcal{L}_{\text{cd}}$ and $\mathcal{L}_{\text{adj}}$ across the full tooth range (1–6), as well as reduced penetration for posterior teeth. At the same time, some metrics show similar median values across models while differing in their distributions. Qualitative inspection reveals artifacts in baseline reconstructions that are not fully captured by vertex-based summary statistics highlighting the limitations of quantitative evaluation.

The multi-stage design exposes a clinically relevant trade-off between pose-agnostic shape fidelity and constraint-aware placement. While the GT pose yields the strongest metric values, it is not available in practice. The TPP+TPO stages provide a controllable refinement mechanism that substantially improves adjacent contact alignment while potentially altering occlusal contact patterns when large pose adjustments are required.

Ablation studies confirm that several domain-specific design choices contribute to stable reconstruction behavior and surface quality, including structured anatomical priors, the CPM, the progressive expansion module, and consistent data normalization. The curvature-weighted chamfer loss introduced to better preserve fine anatomical structures does not yield quantitative improvements, suggesting that architectural and representation level factors dominate the preservation of high-frequency detail in the present setting.

Finally, the end-to-end runtime analysis suggests practical feasibility, with an average per-tooth runtime of approximately 6 s on a single NVIDIA RTX 4090 GPU. This supports a staged workflow, in which a morphological proposal is generated first and contact-driven optimization is applied on demand.

Conclusions: The proposed TST pipeline reconstructs anatomically plausible teeth from segmented 3D scan data while incorporating multi-tooth context and enabling contact-aware placement refinement. By separating shape generation from pose optimization, it provides a flexible framework that addresses both morphological fidelity and functional contact requirements. The quantitative and qualitative results validate the core design principles and demonstrate significant improvements over common PCD completion baselines. Future work should focus on integrating preparation geometry, enhancing high-frequency anatomical detail through architectural changes, and validating robustness across diverse clinical datasets.

Bibliography

[1] Y.-W. Chen, K. Stanley, and A. Wael. "Artificial intelligence in dentistry: current applications and future perspectives". In: *Quintessence Int* 51 (2020), p. 248.

[2] F. Carrillo-Perez et al. "Applications of artificial intelligence in dentistry: A comprehensive review". In: *Journal of Esthetic and Restorative Dentistry* 34.1 (2022), pp. 259–280. DOI: 10.1111/jerd.12844.

[3] H. Ding, J. Wu, W. Zhao, J. P. Matinlinna, M. F. Burrow, and J. K. H. Tsoi. "Artificial intelligence in dentistry—A review". In: *Frontiers in Dental Medicine* 4 (2023). DOI: 10.3389/fdmed.2023.1085251.

[4] F. Schwendicke, W. Samek, and J. Krois. "Artificial Intelligence in Dentistry: Chances and Challenges". In: *Journal of Dental Research* 99.7 (2020), pp. 769–774. DOI: 10.1177/0022034520915714.

[5] Z. Ma, C. Du, Q. Lao, and X. Xie. "Generative Artificial Intelligence: Applications and Future Prospects in Dentistry". In: *International dental journal* 76.1 (2025), p. 108276. DOI: 10.1016/j.identj.2025.108276.

[6] F. Villena, C. Véliz, R. García-Huidobro, and S. Aguayo. "Generative artificial intelligence in dentistry: A narrative review of current approaches and future challenges". In: *Dentistry Review* 5.4 (2025), p. 100160. DOI: 10.1016/j.dentre.2025.100160.

[7] M. Revilla-León et al. "Artificial intelligence applications in restorative dentistry: A systematic review". In: *The Journal of prosthetic dentistry* 128.5 (2022), pp. 867–875. DOI: 10.1016/j.prosdent.2021.02.010.

[8] S. A. Bernauer, N. U. Zitzmann, and T. Joda. "The Use and Performance of Artificial Intelligence in Prosthodontics: A Systematic Review". In: *Sensors (Basel, Switzerland)* 21.19 (2021). DOI: 10.3390/s21196628.

[9] M. Revilla-León et al. "Artificial intelligence models for tooth-supported fixed and removable prosthodontics: A systematic review". In: *The Journal of prosthetic dentistry* 129.2 (2023), pp. 276–292. DOI: 10.1016/j.prosdent.2021.06.001.

[10] L. Lepidi, F. Grande, G. Baldassarre, C. Suriano, J. Li, and S. Catapano. "Preliminary clinical study of the accuracy of a digital axiographic recording system for the assessment of sagittal condylar inclination". In: *Journal of dentistry* 135 (2023), p. 104583. DOI: 10.1016/j.jdent.2023.104583.

[11] Z. Nagy, A. Mikolicz, and J. Vag. "In-vitro accuracy of a novel jaw-tracking technology". In: *Journal of dentistry* 138 (2023), p. 104730. DOI: 10.1016/j.jdent.2023.104730.

[12] A. Broll, S. Hahnel, M. Goldhacker, J. Rossel, M. Schmidt, and M. Rosentritt. "Influence of digital crown design software on morphology, occlusal characteristics, fracture force and marginal fit". In: *Dental Materials* 42.1 (2025), pp. 8–15. DOI: 10.1016/j.dental.2025.09.003.

[13] V. Preis, T. Dowerk, M. Behr, C. Kolbeck, and M. Rosentritt. "Influence of cusp inclination and curvature on the in vitro failure and fracture resistance of veneered zirconia crowns". In: *Clinical Oral Investigations* 18.3 (2014), pp. 891–900. DOI: 10.1007/s00784-013-1029-9.

[14] K. Schnitzhofer, A. Rauch, M. Schmidt, and M. Rosentritt. "Impact of the occlusal contact pattern and occlusal adjustment on the wear and stability of crowns". In: *Journal of dentistry* 128 (2023), p. 104364. DOI: 10.1016/j.jdent.2022.104364.

[15] A. Broll, M. Goldhacker, S. Hahnel, and M. Rosentritt. "Generative deep learning approaches for the design of dental restorations: A narrative review". In: *Journal of dentistry* 145 (2024), p. 104988. DOI: 10.1016/j.jdent.2024.104988.

[16] J.-J. Hwang, S. Azernikov, A. A. Efros, and S. X. Yu. *Learning Beyond Human Expertise with Generative Models for Dental Restorations*. 2018. arXiv: 1804.00064v1.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. DOI: 10.1109/cvpr.2017.632.

[18] F. Yuan et al. "Personalized design technique for the dental occlusal surface based on conditional generative adversarial networks". In: *International Journal for Numerical Methods in Biomedical Engineering* 36.5 (2020), e3321. DOI: 10.1002/cnm.3321.

[19] S. Tian et al. "Efficient Computer-Aided Design of Dental Inlay Restoration: A Deep Adversarial Framework". In: *IEEE Transactions on Medical Imaging* 40.9 (2021), pp. 2415–2427. DOI: 10.1109/tmi.2021.3077334.

[20] S. Tian et al. “A Dual Discriminator Adversarial Learning Approach for Dental Occlusal Surface Reconstruction”. In: *Journal of Healthcare Engineering* 2022 (2022), pp. 1–14. DOI: 10.1155/2022/1933617.

[21] S. Tian et al. “DCPR-GAN: Dental Crown Prosthesis Restoration Using Two-Stage Generative Adversarial Networks”. In: *IEEE Journal of Biomedical and Health Informatics* 26.1 (2022), pp. 151–160. DOI: 10.1109/jbhi.2021.3119394.

[22] Z. Gu, Z. Wu, and N. Dai. “Image generation technology for functional occlusal pits and fissures based on a conditional generative adversarial network”. In: *PloS one* 18.9 (2023), e0291728. DOI: 10.1371/journal.pone.0291728.

[23] A. Broll, M. Rosentritt, T. Schlegl, and M. Goldhacker. “A data-driven approach for the partial reconstruction of individual human molar teeth using generative deep learning”. In: *Frontiers in Artificial Intelligence* 7 (2024), p. 1339193. DOI: 10.3389/frai.2024.1339193.

[24] J. Roh, J. Kim, and J. Lee. “Two-stage deep learning framework for occlusal crown depth image generation”. In: *Computers in biology and medicine* 183 (2024), p. 109220. DOI: 10.1016/j.compbiomed.2024.109220.

[25] J. Z. Ye, T. Ørkild, P. L. Søndergaard, and S. Hauberg. *Variational Autoencoding of Dental Point Clouds*. 2023-07-20. arXiv.org: 2307.10895v4. URL: http://arxiv.org/pdf/2307.10895.

[26] C. Chanintonsongkhla, V. Chouvatut, C. Bunkhumpornpat, and P. Theerasopon. “A latent variable deep generative model for 3D anterior tooth shape”. In: *Journal of Prosthodontics* (2025). DOI: 10.1111/jopr.14092.

[27] R. C. W. Chau, R. T.-C. Hsung, C. McGrath, E. H. N. Pow, and W. Y. H. Lam. “Accuracy of artificial intelligence-designed single-molar dental prostheses: A feasibility study”. In: *The Journal of Prosthetic Dentistry* (2023). DOI: 10.1016/j.prosdent.2022.12.004.

[28] H. Ding et al. “Morphology and mechanical performance of dental crown designed by 3D-DCGAN”. In: *Dental Materials* 39.3 (2023), pp. 320–332. DOI: 10.1016/j.dental.2023.02.001.

[29] O. Lessard, F. Guibault, J. Keren, and F. Cheriet. “Dental Restoration using a Multi-Resolution Deep Learning Approach”. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 3/28/2022 - 3/31/2022, pp. 1–4. DOI: 10.1109/ISBI52829.2022.9761622.

[30] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou. "PoinTr: Diverse Point Cloud Completion with Geometry-Aware Transformers". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. DOI: 10.1109/iccv48922.2021.01227.

[31] X. Yu, Y. Rao, Z. Wang, J. Lu, and J. Zhou. *AdaPoinTr: Diverse Point Cloud Completion with Adaptive Geometry-Aware Transformers*. 2023. DOI: 10.48550/arXiv.2301.04545. arXiv: 2301.04545v1.

[32] H. Zhu, X. Jia, C. Zhang, and T. Liu. "ToothCR: A Two-Stage Completion and Reconstruction Approach on 3D Dental Model". In: *26th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. [S.l.]: Springer, 2022, pp. 161–172. DOI: 10.1007/978-3-031-05981-0_13.

[33] X. Ma, G. Lin, S. Zhao, X. Zou, H. Xie, and M. Tang. "Feasibility and accuracy assessment of AI-driven incisor restoration using 3D point cloud data". In: *Clinical Oral Investigations* 29.11 (2025), p. 532. DOI: 10.1007/s00784-025-06618-5.

[34] G. Hosseinimanesh, F. Ghadiri, F. Guibault, F. Cheriet, and J. Keren. "From Mesh Completion to AI Designed Crown". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part IX*. Ed. by H. Greenspan et al. Vol. 14228. Lecture Notes in Computer Science. Cham: Springer Nature, 2023, pp. 555–565. DOI: 10.1007/978-3-031-43996-4_53.

[35] G. Hosseinimanesh, A. Alsheghri, J. Keren, F. Cheriet, and F. Guibault. "Personalized dental crown design: A point-to-mesh completion network". In: *Medical Image Analysis* 101 (2025), p. 103439. DOI: 10.1016/j.media.2024.103439.

[36] S. Yang et al. "DCrownFormer: Morphology-Aware Point-to-Mesh Generation Transformer for Dental Crown Prosthesis from 3D Scan Data of Antagonist and Preparation Teeth". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by M. G. Linguraru et al. Vol. 15006. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2024, pp. 109–119. DOI: 10.1007/978-3-031-72089-5_11.

[37] S. Yang et al. "DCrownFormer+: Morphology-aware mesh generation and refinement transformer for dental crown prosthesis from 3D scan data of preparation and antagonist teeth". In: *Medical Image Analysis* 105 (2025), p. 103717. DOI: 10.1016/j.media.2025.103717.

[38] Y. Ping, G. Wei, L. Yang, Z. Cui, and W. Wang. "Self-attention implicit function networks for 3D dental data completion". In: *Computer Aided Geometric Design* 90 (2021), p. 102026. DOI: 10.1016/j.cagd.2021.102026.

[39] O. Saleh et al. "Feasibility of using two generative AI models for teeth reconstruction". In: *Journal of dentistry* 151 (2024), p. 105410. DOI: 10.1016/j.jdent.2024.105410.

[40] J. Wu, Y. Huang, J. He, K. Chen, W. Wang, and X. Li. "Automatic restoration and reconstruction of defective tooth based on deep learning technology". In: *BMC oral health* 25.1 (2025), p. 1292. DOI: 10.1186/s12903-025-06576-0.

[41] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. "Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images". In: *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2018, pp. 55–71. DOI: 10.1007/978-3-030-01252-6_4.

[42] M. Liu, X. Li, J. Liu, W. Liu, and Z. Yu. "TUCNet: A channel and spatial attention-based graph convolutional network for teeth upsampling and completion". In: *Computers in Biology and Medicine* 166 (2023), p. 107519. DOI: 10.1016/j.compbiomed.2023.107519.

[43] Du Chen, M.-Q. Yu, Q.-J. Li, X. He, F. Liu, and J.-F. Shen. "Precise tooth design using deep learning-based templates". In: *Journal of dentistry* 144 (2024), p. 104971. DOI: 10.1016/j.jdent.2024.104971.

[44] Y. Feng et al. "3D reconstruction for maxillary anterior tooth crown based on shape and pose estimation networks". In: *International Journal of Computer Assisted Radiology and Surgery* 18.8 (2023), pp. 1405–1416. DOI: 10.1007/s11548-023-02841-1.

[45] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. *PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space*. 2017-06-08. arXiv.org: 1706.02413v1. URL: http://arxiv.org/pdf/1706.02413.

[46] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. DOI: 10.1109/cvpr.2017.16.

[47] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert. "PCN: Point Completion Network". In: *2018 International Conference on 3D Vision (3DV 2018): Verona, Italy, 5-8 September 2018*. Piscataway, NJ: IEEE, 2018, pp. 728–737. DOI: 10.1109/3DV.2018.00088.

[48] Y. Yang, C. Feng, Y. Shen, and D. Tian. *FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation*. 2017. arXiv.org: 1712.07262v2. URL: http://arxiv.org/pdf/1712.07262v2.

[49] S. Li, P. Gao, X. Tan, and M. Wei. "ProxyFormer: Proxy Alignment Assisted Point Cloud Completion with Missing Part Sensitive Transformer". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 9466–9475. DOI: 10.1109/CVPR52729.2023.00913.

[50] H. Zhou et al. "SeedFormer: Patch Seeds Based Point Cloud Completion with Upsample Transformer". In: *Computer vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022 : proceedings*. Ed. by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner. Vol. 13663. Lecture Notes in Computer Science. Cham: Springer, 2022, pp. 416–432. DOI: 10.1007/978-3-031-20062-5_24.

[51] J. Liu and X. Wang. "PoinTr-PM: Diverse Point Cloud Completion with Geometry-Aware Transformers and Point Moving". In: *2024 43rd Chinese Control Conference (CCC)*. IEEE, 7/28/2024 - 7/31/2024, pp. 4381–4386. DOI: 10.23919/CCC63176.2024.10662066.

[52] X. Wen et al. "PMP-Net++: Point Cloud Completion by Transformer-Enhanced Multi-Step Point Moving Paths". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2023), pp. 852–867. DOI: 10.1109/TPAMI.2022.3159003.

[53] X. Yan et al. "FBNet: Feedback Network for Point Cloud Completion". In: *Computer Vision - ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part II*. Ed. by S. Avidan. Springer, 2022, pp. 676–693. DOI: 10.1007/978-3-031-20086-1_39.

[54] Y. Rong, H. Zhou, L. Yuan, C. Mei, J. Wang, and T. Lu. "CRA-PCN: Point Cloud Completion with Intra- and Inter-level Cross-Resolution Transformers". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.5 (2024), pp. 4676–4685. DOI: 10.1609/aaai.v38i5.28268.

[55] X. Yan, L. Lin, N. J. Mitra, D. Lischinski, D. Cohen-Or, and H. Huang. "ShapeFormer: Transformer-based Shape Completion via Sparse Representation". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Ed. by I. o. E. Engineers and Electronics. Piscataway, NJ: IEEE, 2022, pp. 6229–6239. DOI: 10.1109/CVPR52688.2022.00614.

[56] J. Wang, Y. Cui, D. Guo, J. Li, Q. Liu, and C. Shen. "PointAttN: You Only Need Attention for Point Cloud Completion". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.6 (2024), pp. 5472–5480. DOI: 10.1609/aaai.v38i6.28356.

[57] S. Mo et al. *DiT-3D: Exploring Plain Diffusion Transformers for 3D Shape Generation*. 2023-07-04. arXiv.org: 2307.01831v1. URL: http://arxiv.org/pdf/2307.01831.

[58] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. *Dynamic Graph CNN for Learning on Point Clouds*. 2018. arXiv: 1801.07829v2.

[59] J. Huang, Z. Gojcic, M. Atzmon, O. Litany, S. Fidler, and F. Williams. *Neural Kernel Surface Reconstruction*. 2023-05-31. arXiv.org: 2305.19590v2. URL: http://arxiv.org/pdf/2305.19590.

[60] A. Vaswani et al. *Attention Is All You Need*. 2017-06-12. arXiv.org: 1706.03762v7. URL: http://arxiv.org/pdf/1706.03762.

[61] J. L. Ba, J. R. Kiros, and G. E. Hinton. *Layer Normalization*. 2016-07-21. arXiv.org: 1607.06450v1. URL: https://arxiv.org/pdf/1607.06450.

[62] S. Wang et al. "A 3D dental model dataset with pre/post-orthodontic treatment for automatic tooth alignment". In: *Scientific data* 11.1 (2024), p. 1277. DOI: 10.1038/s41597-024-04138-7.

[63] D. Cohen-Steiner and J.-M. Morvan. "Restricted delaunay triangulations and normal cycle". In: *Proceedings of the nineteenth conference on Computational geometry - SCG '03*. New York, New York, USA: ACM Press, 2003. DOI: 10.1145/777837.777839.

[64] A. Broll, M. Goldhacker, S. Hahnel, and M. Rosentritt. "Morphological effects of input data quantity in AI-powered dental crown design". In: *Journal of dentistry* (2025), p. 105767. DOI: 10.1016/j.jdent.2025.105767.

[65] S. Qiu, S. Anwar, and N. Barnes. "PU-Transformer: Point Cloud Upsampling Transformer". In: *Lecture Notes in Computer Science*. Cham: Springer Nature Switzerland, 2023, pp. 326–343. DOI: 10.1007/978-3-031-26319-4_20.

[66] R. Li, X. Li, P.-A. Heng, and C.-W. Fu. "Point Cloud Upsampling via Disentangled Refinement". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 344–353. DOI: 10.1109/cvpr46437.2021.00041.

[67] H. Fan, H. Su, and L. Guibas. "A Point Set Generation Network for 3D Object Reconstruction from a Single Image". In: *30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ: IEEE, 2017, pp. 2463–2471. DOI: 10.1109/CVPR.2017.264.

[68] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778. DOI: 10.1109/cvpr.2016.90.

[69] T. Kim. *Generalizing MLPs With Dropouts, Batch Normalization, and Skip Connections*. 2021-08-18. arXiv.org: 2108.08186v2. URL: https://arxiv.org/pdf/2108.08186.

[70] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. "The ball-pivoting algorithm for surface reconstruction". In: *IEEE Transactions on Visualization and Computer Graphics* 5.4 (1999), pp. 349–359. DOI: 10.1109/2945.817351.

[71] M. Kazhdan, M. Bolitho, and H. Hoppe. "Poisson Surface Reconstruction". In: *1727-8384* (2006).

[72] D. Comaniciu and P. Meer. "Mean shift: a robust approach toward feature space analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002), pp. 603–619. DOI: 10.1109/34.1000236.

[73] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. 2017. arXiv.org: 1711.05101. URL: https://arxiv.org/pdf/1711.05101.

[74] F. Lin et al. "InfoCD: A Contrastive Chamfer Distance Loss for Point Cloud Completion". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc, 2023, pp. 76960–76973.

A Appendix

A.1 Literature research

This section documents the literature research workflow used to compile a focused reference corpus of DL-based methods for dental reconstruction from geometric scan data. The search targeted approaches that operate directly on scan-derived surface geometry, in particular meshes and PCDs, and that aim to reconstruct or generate dental anatomy for restorative workflows. Studies relying primarily on radiographic or tomographic imaging modalities (e.g., computed tomography, cone-beam computed tomography, magnetic resonance imaging, or X-ray radiography) were excluded to keep the scope aligned with CAD-oriented crown and restoration design. The literature research was last updated on 01.11.2025. For PubMed, the final query was restricted to publication years 2016–2026.

PubMed served as the primary biomedical database, complemented by Google Scholar to capture relevant conference proceedings and technical venues that are not consistently indexed in PubMed. The PubMed search query was refined iteratively, combining three term groups:

- dental anatomy and prosthodontic concepts (e.g., tooth, crown, occlusal surface, prosthesis, restoration),
- geometric representations (e.g., PCDs, meshes, 3D scans, dental models), and
- DL and generative modeling terminology (e.g., GANs, transformers, autoencoders, diffusion models, implicit functions).

The final refined PubMed query returned 188 records. To ensure reproducibility, the complete query string is reported at the end of this section.

The resulting curated corpus comprises 25 publications on DL-based dental reconstruction from meshes and scanned surface representations. Of these, 17 papers were directly retrievable from PubMed via the final refined query. The remaining 8 publications were not indexed in PubMed and were therefore recovered by (i) cross-references within the

PubMed-retrieved papers and (ii) targeted Google Scholar searches using distinctive title phrases and model names.

The literature research confirmed that PubMed alone does not provide comprehensive coverage of the relevant technical literature for geometry-based dental reconstruction. Consequently, the combination of database search, citation chasing, and targeted Google Scholar retrieval was required to obtain a corpus that spans both clinically oriented journal work and technically focused computer-vision contributions.

The literature research was intentionally scoped to PubMed and Google Scholar and therefore inherits their coverage, indexing, and ranking biases. The PubMed query used title/abstract field constraints and a predefined inclusion/exclusion logic. Relevant publications that do not use the selected terminology, are not indexed, or fall outside the chosen time window may have been missed. Finally, the workflow was designed to compile a focused state-of-the-art corpus rather than to perform a formal systematic review. Therefore, no systematic-review flow diagram is reported. The following search strategy was employed for the literature review:

```
(
  (
    dental[tiab] OR tooth[tiab] OR teeth[tiab] OR molar[tiab] OR occlusal[tiab]
  )
  AND
  (
    "point cloud"[tiab]
    OR mesh[tiab]
    OR "mesh completion"[tiab]
    OR "point-to-mesh"[tiab]
    OR "implicit function"[tiab]
    OR "graph convolutional"[tiab]
    OR "depth image"[tiab]
    OR "3D scan"[tiab]
    OR "3D model"[tiab]
    OR "3D dental model"[tiab]
    OR "dental model"[tiab]
    OR "shape completion"[tiab]
    OR "surface reconstruction"[tiab]
  )
  AND
  (
```

```
    "deep learning"[tiab]
    OR neural network*[tiab]
    OR generative[tiab]
    OR adversarial[tiab]
    OR GAN[tiab]
    OR StyleGAN[tiab]
    OR transformer[tiab]
    OR diffusion[tiab]
    OR "artificial intelligence"[tiab]
    OR "machine learning"[tiab]
    OR PointFlow[tiab]
    OR "self-attention"[tiab]
    OR "self attention"[tiab]
    OR "variational autoencoder"[tiab]
    OR autoencoder*[tiab]
    OR autoencoding[tiab]
  )
)
OR
(
  (
    "dental crown"[tiab]
    OR "tooth crown"[tiab]
    OR "dental prosthesis"[tiab]
    OR "dental prostheses"[tiab]
    OR "dental restoration"[tiab]
    OR "dental restorations"[tiab]
    OR inlay[tiab]
    OR "inlay restoration"[tiab]
    OR "tooth shape"[tiab]
    OR "tooth design"[tiab]
    OR "crown design"[tiab]
    OR "crown prosthesis"[tiab]
    OR "occlusal surface"[tiab]
    OR "occlusal pits"[tiab]
    OR "occlusal fissures"[tiab]
    OR "anterior tooth"[tiab]
  )
  AND
  (
    "deep learning"[tiab]
    OR neural network*[tiab]
```

```
    OR generative[tiab]
    OR adversarial[tiab]
    OR GAN[tiab]
    OR StyleGAN[tiab]
    OR transformer[tiab]
    OR diffusion[tiab]
    OR "artificial intelligence"[tiab]
    OR "machine learning"[tiab]
    OR PointFlow[tiab]
    OR "self-attention"[tiab]
    OR "self attention"[tiab]
    OR "variational autoencoder"[tiab]
    OR autoencoder*[tiab]
    OR autoencoding[tiab]
  )
)
NOT
(
  radiograph*[tiab]
  OR radiography[tiab]
  OR radiographic[tiab]
  OR "cone beam"[tiab]
  OR "cone-beam"[tiab]
  OR CBCT[tiab]
  OR "X-ray"[tiab]
  OR "X ray"[tiab]
  OR tomography[tiab]
  OR "computed tomography"[tiab]
  OR CT[tiab]
  OR MRI[tiab]
)
```

A.2 Per tooth quantitative results

The following tables (Table A.1, Table A.2, Table A.3) contain the complete quantitative results for the model comparison as well as the data context and pose ablations discussed in Chapter 3 with per tooth breakdowns of the median test-set metrics. Aligning with previous sections, units are omitted in the following tables for brevity. The units and short metric descriptors are provided in Table 2.2.

Table A.1: Median test-set metrics for TST with TPO, PoinTr, and PCN per tooth. Lower is better.

Configuration		Evaluation metrics (median)						
Model	Tooth	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	$\mathcal{L}_{\text{cp,num}}$	$\mathcal{L}_{\text{adj}}$
TST	1	**0.219**	0.341	0.000	0.000	0.000	0.000	**0.067**
PoinTr	1	0.321	0.340	0.000	0.000	0.000	0.000	0.455
PCN	1	0.280	**0.323**	0.000	0.000	0.000	0.000	0.294
TST	2	**0.250**	**0.359**	0.006	0.178	0.000	0.000	**0.056**
PoinTr	2	0.360	0.377	0.003	0.187	0.000	0.000	0.503
PCN	2	0.313	0.365	**0.000**	**0.000**	0.000	0.000	0.348
TST	3	**0.295**	0.458	0.052	**0.527**	0.000	0.000	**0.064**
PoinTr	3	0.388	0.423	0.074	0.800	0.000	0.000	0.479
PCN	3	0.345	**0.406**	**0.039**	0.557	0.000	0.000	0.358
TST	4	**0.271**	0.318	**0.039**	**1.150**	**1.345**	1.000	**0.062**
PoinTr	4	0.345	0.304	0.070	1.411	2.033	1.000	0.515
PCN	4	0.331	**0.291**	0.055	1.166	2.061	1.000	0.297
TST	5	**0.245**	0.263	**0.042**	1.034	1.240	1.000	**0.058**
PoinTr	5	0.325	0.269	0.069	1.024	**0.836**	1.000	0.400
PCN	5	0.309	**0.247**	0.054	**0.856**	0.944	1.000	0.247
TST	6	**0.293**	**0.308**	0.044	1.261	1.193	2.000	**0.085**
PoinTr	6	0.348	0.346	0.049	1.231	1.102	1.000	0.433
PCN	6	0.318	0.319	0.049	**1.193**	**0.764**	1.000	0.300

Table A.2: Median test-set metrics for TST across pose stages t_{GT}, t_{pred}, and t_{opt} per tooth. Lower is better.

Configuration		Evaluation metrics (median)						
Pose	Tooth	$\mathcal{L}_{\mathrm{cd}}$	$\mathcal{L}_{\mathrm{cIoU}}$	$\mathcal{L}_{\mathrm{pen}}$	$\mathcal{L}_{\mathrm{cp,pos}}$	$\mathcal{L}_{\mathrm{cp,dist}}$	$\mathcal{L}_{\mathrm{cp,num}}$	$\mathcal{L}_{\mathrm{adj}}$
t_{GT}	1	**0.180**	0.338	0.000	0.000	0.000	0.000	0.333
t_{pred}	1	0.215	0.341	0.000	0.000	0.000	0.000	0.418
t_{opt}	1	0.219	0.341	0.000	0.000	0.000	0.000	**0.067**
t_{GT}	2	**0.183**	**0.338**	0.000	0.000	0.000	0.000	0.385
t_{pred}	2	0.237	0.343	0.000	0.000	0.000	0.000	0.433
t_{opt}	2	0.250	0.359	0.006	0.178	0.000	0.000	**0.056**
t_{GT}	3	**0.215**	**0.436**	**0.034**	**0.332**	0.000	0.000	0.410
t_{pred}	3	0.267	0.446	0.049	0.414	0.000	0.000	0.455
t_{opt}	3	0.295	0.458	0.052	0.527	0.000	0.000	**0.064**
t_{GT}	4	**0.210**	**0.296**	0.048	**0.840**	**0.599**	0.000	0.310
t_{pred}	4	0.269	0.308	0.059	1.057	1.515	1.000	0.347
t_{opt}	4	0.271	0.318	**0.039**	1.150	1.345	1.000	**0.062**
t_{GT}	5	**0.196**	**0.248**	0.049	**0.723**	**0.707**	1.000	0.291
t_{pred}	5	0.252	0.260	0.055	1.054	1.407	1.000	0.342
t_{opt}	5	0.245	0.263	**0.042**	1.034	1.240	1.000	**0.058**
t_{GT}	6	**0.226**	**0.302**	**0.041**	**0.953**	**0.599**	1.000	0.325
t_{pred}	6	0.295	0.311	0.064	1.169	0.992	1.000	0.357
t_{opt}	6	0.293	0.308	0.044	1.261	1.193	2.000	**0.085**

Table A.3: Median test-set metrics for TST trained with different context sizes (full, quad, adj), evaluated on pose agnostic prediction output per tooth. Lower is better.

Configuration		Evaluation metrics (median)						
Context	Tooth	$\mathcal{L}_{\text{cd}}$	$\mathcal{L}_{\text{cIoU}}$	$\mathcal{L}_{\text{pen}}$	$\mathcal{L}_{\text{cp,pos}}$	$\mathcal{L}_{\text{cp,dist}}$	$\mathcal{L}_{\text{cp,num}}$	$\mathcal{L}_{\text{adj}}$
full	1	0.180	0.338	0.000	0.000	0.000	0.000	**0.333**
quad	1	0.191	0.347	0.000	0.000	0.000	0.000	0.367
adj	1	**0.170**	**0.328**	0.000	0.000	0.000	0.000	0.350
full	2	0.183	**0.338**	0.000	0.000	0.000	0.000	0.385
quad	2	0.180	0.354	0.000	0.000	0.000	0.000	0.398
adj	2	**0.174**	0.346	0.000	0.000	0.000	0.000	**0.377**
full	3	0.215	0.436	0.034	**0.332**	0.000	0.000	0.410
quad	3	0.210	**0.419**	**0.027**	0.357	0.000	0.000	**0.360**
adj	3	**0.208**	0.433	0.028	0.341	0.000	0.000	0.376
full	4	**0.210**	0.296	0.048	**0.840**	**0.599**	0.000	0.310
quad	4	0.211	**0.282**	**0.041**	0.920	1.135	1.000	**0.308**
adj	4	0.218	0.291	0.046	0.894	0.807	1.000	0.317
full	5	**0.196**	0.248	0.049	**0.723**	0.707	1.000	**0.291**
quad	5	0.208	**0.237**	0.051	0.757	0.732	1.000	0.311
adj	5	0.215	0.265	**0.046**	0.737	**0.583**	1.000	0.309
full	6	**0.226**	0.302	0.041	**0.953**	0.599	1.000	**0.325**
quad	6	0.241	**0.296**	0.045	0.987	**0.540**	1.000	0.348
adj	6	0.234	0.301	**0.040**	1.020	0.816	1.000	0.331

Danksagung

An dieser Stelle möchte ich allen danken, die mich während der Durchführung dieser Arbeit unterstützt und begleitet haben. Mein besonderer Dank gilt an erster Stelle meinem Mentor Prof. Dr. M. Rosentritt. Zunächst einmal wäre ich wohl ohne ihn niemals auf die Idee gekommen, dieses Thema als Applikation von generativem Deep Learning in der Zahnmedizin anzugehen. Seine Begeisterung für das, was ich in dieser Arbeit mache und erreicht habe, sowie sein Engagement und Interesse haben mich stets motiviert und mir geholfen, immer mein Bestes zu geben. Darüber hinaus hat er mich durch seine konstruktiven Rückmeldungen und vielen hilfreichen Diskussionen immer wieder auf neue Ideen gebracht, die diese Arbeit maßgeblich geprägt haben. Außerdem möchte ich meinem Betreuer Prof. Dr. S. Hahnel danken, der die Rahmenbedingungen für diese Arbeit geschaffen hat und es mir ermöglicht hat, in einer so tollen Abteilung zu arbeiten. Vielen Dank an meinen Mentor Prof. Dr. M. Goldhacker, der mich von Anfang an als Betreuer meiner Masterarbeit begleitet hat und mich durch seine Vorlesung überhaupt erst auf das Thema Deep Learning aufmerksam gemacht hat. Schon damals konnte er seine Begeisterung auf mich übertragen und hat somit den Grundstein für das Entstehen dieser Arbeit gelegt.

Ich möchte mich auch bei Prof. Dr. T. Schlegl, dem ehemaligen Leiter des Robotiklabors der OTH Regensburg, bedanken. Wie Prof. Dr. M. Rosentritt hat auch er diese Arbeit überhaupt erst ermöglicht, indem er den Kontakt zum Universitätsklinikum Regensburg hergestellt und meine Masterarbeit in seinem Labor durch einen festen Arbeitsplatz, die benötigte Rechenleistung und so manche interessante Diskussion unterstützt hat. Vielen Dank an die gesamte Prothetik, die mich so herzlich aufgenommen hat und mir ein tolles Arbeitsumfeld geboten hat. Dabei sind nicht nur interessante fachliche Gespräche entstanden, sondern auch so viele lustige und schöne Momente, die ich nicht missen möchte. Alle zusammen haben dazu beigetragen, dass ich mich willkommen und wohl gefühlt habe und die Zeit dort wirklich genießen konnte.

Vielen Dank an meine Partnerin Vera. Du hast die gesamte Reise mit mir zusammen begonnen, beendet und warst immer unterstützend an meiner Seite. Dafür bin ich dir unendlich dankbar.

An letzter Stelle möchte ich meinen Eltern danken. Ohne eure harte Arbeit, den unermüdlichen Einsatz für eure Kinder, eure Unterstützung und euer Vertrauen in mich, wäre ich heute nicht an dieser Stelle. Danke, dass ihr immer für mich da wart.

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater*in oder andere Personen) in Anspruch genommen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Regensburg, 03.02.2026 Alexander Broll