UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

UNIVERSITY OF MODENA AND REGGIO EMILIA

"Enzo Ferrari" Department of Engineering

Master's Degree in Artificial Intelligence Engineering (LM-32)

An Automated Dental Bracket Positioning Model: A Tailored AI Solution for Clinical Orthodontic Practice

Supervisor:
Federico Bolelli

Co-supervisor:
Luca Lumetti

Candidate:
Matteo Lugli
Student ID 196569

ACADEMIC YEAR 2024/2025

Contents

List of Figures

Abstract

An Automated Dental Bracket Positioning Model: A Tailored AI Solution for Clinical Orthodontic Practice

Accurate positioning of orthodontic brackets is a critical step in fixed appliance therapy, as placement errors directly affect treatment efficiency, duration, and clinical outcomes. Recent advances in digital orthodontics and intra-oral scanning enable the automation of this process through data-driven methods. This thesis presents a deep learning framework for predicting dental bracket installation points directly from intra-oral scans of patients.

The proposed approach operates on three-dimensional dental data and leverages point cloud representations to model tooth geometry. A complete pipeline is developed, including data collection, preprocessing, tooth segmentation, and bracket position prediction. Multiple prediction strategies are investigated, including direct regression and heatmap-based formulations, and their performance is evaluated on curated datasets derived from clinical annotations. To support model deployment in real-world scenarios, a dedicated segmentation model is designed and integrated, providing improved flexibility and control compared to off-the-shelf solutions.

Extensive experimental results demonstrate that the proposed methods achieve accurate and consistent bracket placement predictions, approaching clinically acceptable precision. Additionally, the system is packaged and deployed through a web-based API, enabling seamless integration with existing orthodontic software and workflows.

This work highlights the feasibility of automated bracket placement from intra-oral scans and provides a scalable foundation for future improvements through additional expert annotations, model refinement, and extended evaluation of segmentation performance. The proposed framework contributes toward reducing manual effort and variability in orthodontic bracket positioning, supporting more efficient and standardized digital orthodontic treatments.

Keywords: Medical Imaging, Deep Learning, Segmentation, Orthodontics, 3D, Brackets

1. Introduction

Orthodontic treatment planning increasingly relies on digital workflows based on three-dimensional intra-oral scans (IOS) [11, 25]. These scans enable precise visualization of dental anatomy and support computer-aided design (CAD) procedures for appliances such as brackets, aligners, and customized orthodontic devices. Despite significant advances in dental digitization and learning-based tooth segmentation, the placement of orthodontic brackets on 3D models remains largely a manual and time-consuming task performed by trained specialists [2].

In current clinical practice, orthodontists or dental technicians manually identify suitable bracket installation points on each tooth surface using dedicated CAD software. This process requires repeated visual inspection, fine-grained manipulation of 3D geometry, and expert anatomical knowledge. As a result, bracket positioning represents a major bottleneck in the digital orthodontic pipeline, particularly in high-throughput clinical or industrial settings. Reducing the time required for this step without compromising placement accuracy would significantly improve the scalability and efficiency of orthodontic workflows.

While prior research has addressed related subproblems such as tooth segmentation from intra-oral scans [35, 36, 15] and landmark detection [14, 26, 10] on dental surfaces, the specific problem of automatically predicting orthodontic bracket installation points directly from intra-oral scans has not yet been tackled in a complete and production-ready manner. Existing commercial systems still rely on manual intervention, and published academic approaches are often limited to simplified geometries, two-dimensional projections, or heuristic rules that do not generalize well across patients and tooth types.

In this thesis, we address this gap by proposing a learning-based system for automatic orthodontic bracket placement that operates directly on 3D intra-oral scans. The proposed approach formulates bracket positioning as a point-wise prediction problem on tooth-level point clouds and leverages a heatmap-based neural architecture to localize optimal installation points on the facial surface of each tooth. Unlike rule-based or purely geometric methods, the model learns clinically relevant placement patterns from data and adapts to anatomical variability across patients.

A key contribution of this work is not only the development of the model itself, but its integration

into a production-level pipeline. The proposed system is designed to operate downstream of an existing tooth instance segmentation network and to be deployable in real-world orthodontic workflows. By providing accurate bracket placement predictions at inference time, the model substantially reduces the amount of manual interaction required during the 3D modeling phase, thereby decreasing the time burden on specialists.

Experimental results demonstrate that the proposed method achieves lower placement error compared to geometric baselines and landmark-based approaches across both maxillary and mandibular arches. Furthermore, per-tooth analysis shows consistent performance across different tooth types, supporting the robustness of the learned representation. These results indicate that the proposed solution is suitable for practical deployment and represents a meaningful step toward fully automated orthodontic appliance design.

The main contributions of this work are summarized as follows:

1. **Data collection and expert annotation.** We collect a novel dataset in collaboration with a professional orthodontic studio in Ferrara. The dataset consists of intra-oral scans enriched with per-tooth expert annotations describing bracket placement. For each tooth, the annotations include clinically relevant landmarks and the definition of the plane on which the orthodontic bracket lies, expressed in the three-dimensional reference frame of the scan. To the best of our knowledge, this is the first dataset providing detailed, tooth-level bracket placement information derived from real clinical workflows.

2. **Dataset preprocessing and release.** We design and implement a comprehensive preprocessing and cleaning pipeline to transform raw clinical data into a consistent and structured dataset. This process includes geometry normalization, annotation validation, and the resolution of inconsistencies across scans and tooth instances. The resulting dataset is suitable for training and evaluating learning-based models for orthodontic applications and is released with the goal of enabling reproducibility and fostering further research in automatic bracket placement and dental scan analysis.

3. **Lightweight semantic segmentation with weak supervision.** We train and release a lightweight semantic segmentation model for intra-oral scans, specifically designed to be suitable for development and integration scenarios. The model achieves accurate tooth-level

segmentation while maintaining very fast inference times (approximately 1.5 seconds per scan). Importantly, the model is trained using only weak supervision, significantly reducing the annotation burden typically associated with dense segmentation tasks in three-dimensional dental data.

4. **Learning-based bracket positioning via point transformers.** We propose a bracket positioning model that predicts three clinically meaningful landmarks on the tooth surface: gingival, incisal, and bracket points. The model leverages PointTransformerV3 [39], a state-of-the-art architecture for point cloud processing, to accurately capture local geometric features and long-range contextual information. Experimental results demonstrate that the proposed approach achieves precise and robust landmark localization across different tooth types.

5. **Production deployment and API integration.** We put the complete pipeline into production and make it accessible through an application programming interface (API). This design choice allows the proposed system to be easily integrated into existing orthodontic software and digital workflows, effectively bridging the gap between academic research and real-world clinical usage.

By addressing a previously unexplored yet clinically relevant problem, and by jointly tackling task definition, data acquisition, and model design, this work contributes to the broader goal of automating digital orthodontic workflows and improving the efficiency and scalability of modern dental care.

2. Intra-oral Scans and Dental Bracket Installation

2.1 Introduction

The increasing digitalization of orthodontic workflows has led to the widespread adoption of three-dimensional (3D) data for diagnosis, treatment planning, and appliance placement. In this context, intra-oral scans have become a primary source of geometric information, providing high-resolution digital representations of a patient's dentition. These scans enable detailed analysis of tooth morphology and spatial relationships, which are essential for a variety of orthodontic procedures.

One of the most critical applications of intra-oral scans in fixed appliance therapy is orthodontic bracket placement [25]. The final position of each bracket directly influences the direction and magnitude of orthodontic forces and, consequently, the effectiveness and duration of treatment. Accurate bracket positioning therefore requires millimeter-level precision and relies on the identification of anatomically meaningful reference points on the tooth surface, commonly referred to as dental landmarks. These landmarks provide the geometric basis for defining bracket positions in both traditional and digital orthodontic workflows.

In this work, intra-oral scans constitute the fundamental input data for an automated system aimed at predicting orthodontic bracket installation points directly from tooth geometry. This chapter introduces the clinical and technical foundations underlying this approach. First, the role of intra-oral scans in orthodontics is discussed, highlighting their relevance for digital bracket placement. Subsequently, the concept of dental landmarks and the measurement-based procedures used in clinical bracket bonding are described, establishing the principles that motivate and guide the proposed automation pipeline.

2.2 Intra-oral Scans in Orthodontics

Intra-oral scans are digital three-dimensional representations of the oral cavity acquired using optical scanning devices known as intra-oral scanners. These devices capture the visible surfaces of teeth and surrounding anatomical structures by projecting light patterns and reconstructing surface

geometry through optical sensing and computational algorithms. The resulting scan provides a high-resolution surface model of the dentition, typically covering individual tooth crowns and, in some cases, portions of the gingiva. Figure 2.1 shows the resulting 3D digital models of the maxillary and mandibular arches, displayed from anterior and lateral perspectives to illustrate the captured surface geometry.

In orthodontics, intra-oral scans have largely replaced conventional physical impressions due to their improved accuracy, efficiency, and patient comfort. Digital scans eliminate the need for impression materials, reduce acquisition time, and allow immediate visualization and analysis of the patient's dentition. Moreover, they can be easily stored, duplicated, and integrated into digital workflows, making them well suited for computer-aided orthodontic applications.

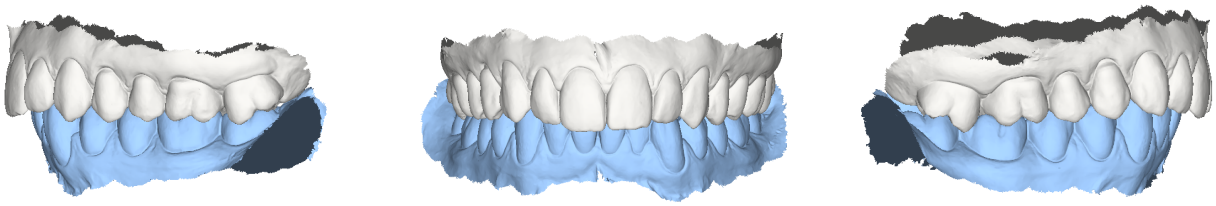
A key application of intra-oral scans in orthodontics is the planning and execution of fixed appliance treatment, particularly the placement of brackets on tooth surfaces. Bracket positioning is a critical step, as the final alignment of teeth depends heavily on the initial placement accuracy. Traditionally, this process relies on the clinician's experience and manual judgment, which can lead to variability across practitioners and cases.

By providing a precise 3D representation of each tooth, intra-oral scans enable bracket placement to be analyzed and planned digitally. Clinicians can examine tooth orientation, curvature, and anatomical landmarks in a virtual environment, supporting more consistent and reproducible placement decisions. Furthermore, the availability of large collections of intra-oral scans [3] makes it possible to apply data-driven methods, such as machine learning, to assist or partially automate this process.

In the context of this thesis, intra-oral scans serve as the geometric foundation for an automated pipeline that predicts bracket installation points directly from tooth morphology. Their high resolution and geometric fidelity make them particularly suitable for learning-based approaches that operate on 3D data.

2.3 Dental Landmarks

Dental landmarks [37] are anatomically defined reference points located on the surface of a tooth crown. They correspond to salient morphological features that are consistently identifiable across



(a) Left (b) Front (c) Right

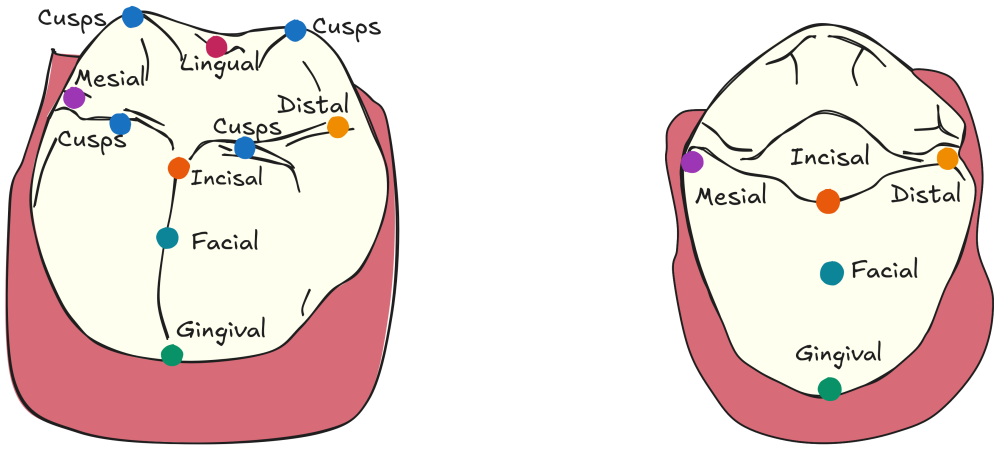
Figure 2.1: Intra-oral scan of a sample case illustrating the mandibular (blue) and maxillary (white) arches. Views are provided from patient's left (a), anterior (b), and right (c) lateral perspectives.

patients and are relevant for clinical assessment and orthodontic treatment planning. Typical examples include incisal edges, cusp tips, and characteristic extrema on the facial, lingual, mesial, or distal aspects of the tooth crown. Figure 2.2 illustrates representative landmarks for molar and premolar teeth, highlighting the differences in crown morphology and available reference features across tooth types.

In orthodontics, dental landmarks serve as stable geometric anchors from which measurements can be taken and spatial relationships can be defined. Their reproducibility is essential, as many clinical procedures rely on millimeter-scale distances measured relative to these points. In the context of bracket placement, landmarks are used to determine vertical positioning along the long axis of the tooth, as well as mesio-distal alignment and rotational orientation.

The definition and availability of landmarks vary significantly depending on tooth anatomy. Posterior teeth, such as molars and premolars, typically present well-defined cusps that can be used as reliable occlusal reference points. In contrast, anterior teeth, particularly incisors, lack pronounced cusps and instead rely on incisal edges or crown contours as primary landmarks.

In a digital setting, dental landmarks are represented as three-dimensional coordinates on a surface mesh or point cloud derived from an intra-oral scan. This representation allows landmarks to be used directly in geometric computations, such as distance measurement, surface normal estimation, and coordinate transformations. From a computational perspective, the reliable identification of such landmarks is a prerequisite for automating measurement-based procedures, including the determination of orthodontic bracket placement.



(a) Molar

(b) Premolar

Figure 2.2: Comparison of dental landmarks for molar (a) and premolar (b) teeth. Note the absence of defined cusps on the premolar.

2.4 The Bonding Procedure

Orthodontic bracket placement relies on the use of dental landmarks to determine both the vertical and horizontal positioning of the bracket on the tooth crown. In standard clinical practice, the vertical position of the bracket is defined as a fixed distance, measured in millimeters, from an anatomically meaningful landmark located at the occlusal or incisal extremity of the tooth. This distance is prescribed by orthodontic guidelines and varies according to tooth type, dental arch, and treatment objectives.

In modern digital workflows, this procedure is carried out on a three-dimensional model of the dentition derived from an intra-oral scan. For each tooth, the orthodontic specialist performs a sequence of manual operations that replicate the traditional measurement-based bonding protocol within a virtual environment. These operations are illustrated in Figure 2.3 and can be summarized as follows:

(a) **Manual selection of anatomical landmarks and estimation of the tooth axis**: the clinician manually identifies two key reference points on the tooth surface: the gingival point and the incisal point. The specific landmarks selected depend on tooth category (incisors, canines, premolars, or molars), reflecting differences in crown morphology. These points are used

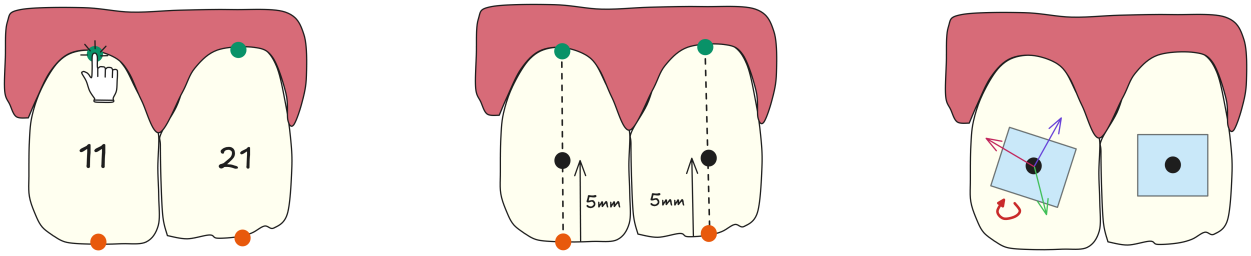
to estimate the principal (long) axis of the tooth, which serves as a geometric reference for subsequent measurements.

(b) **Selection of the bracket installation point**: starting from the vestibular or incisal landmark, the clinician determines the target bracket position by measuring a distance along the tooth surface toward the gingival landmark. This distance is defined in millimeters and depends on the tooth's FDI index, illustrated in Figure 2.4, following standardized orthodontic prescriptions. Importantly, this measurement is most appropriately defined as a geodesic distance on the tooth surface, as it follows the natural curvature of the crown and therefore respects the underlying tooth morphology. The use of geodesic distance ensures that the prescribed measurement remains consistent across teeth with different shapes and surface curvatures. The recommended values for each tooth type are summarized in Figure 2.5. In practice, this step often requires multiple manual interactions, as many 3D modeling software tools used in clinical settings do not provide native support for geodesic distance computation on curved surfaces. As a result, clinicians approximate the desired distance through iterative measurements and adjustments.

(c) **Manual placement and adjustment of the bracket support plane**: once the target installation point has been determined, the clinician manually defines a plane onto which the three-dimensional model of the orthodontic bracket is placed. This plane is oriented and adjusted to conform to the local tooth surface geometry, ensuring appropriate contact, alignment, and rotational orientation of the bracket relative to the tooth crown.

As illustrated in Figure 2.5, the recommended vertical bracket heights differ across tooth categories and may be adjusted to accommodate individual anatomical variability or specific clinical goals. Reported values typically include mean reference heights (e.g., 5.0 mm for maxillary central incisors and canines), along with permissible offsets such as increases of up to +1.5 mm or reductions of up to -1.0 mm.

Because these measurements are performed with millimeter-level precision, accurate identification of dental landmarks is a prerequisite for correct bracket placement. Errors in landmark localization directly propagate to errors in bracket positioning, potentially affecting the overall treatment outcome. While the transition from physical instruments to digital environments has improved



(a) Selection (b) Measurement (c) Plane adjustment

Figure 2.3: Selection (a), measurement (b), and plane adjustment (c) procedures performed by an orthodontic specialist on the patient's intra-oral scan for each individual tooth, illustrating the manual workflow used to determine clinically relevant reference geometry.

visualization and reproducibility, the procedure remains conceptually identical and continues to rely heavily on manual input and clinician expertise. Figure 2.6 shows a screenshot of the annotation software used by an orthodontic specialist for manual bracket point selection.

The explicit reliance of this workflow on anatomically defined landmarks and measurement-based rules makes orthodontic bracket placement particularly well suited for automation. If dental landmarks can be reliably predicted from intra-oral scans, the subsequent determination of bracket installation points and orientations can be performed algorithmically, reducing manual effort while preserving established clinical principles.

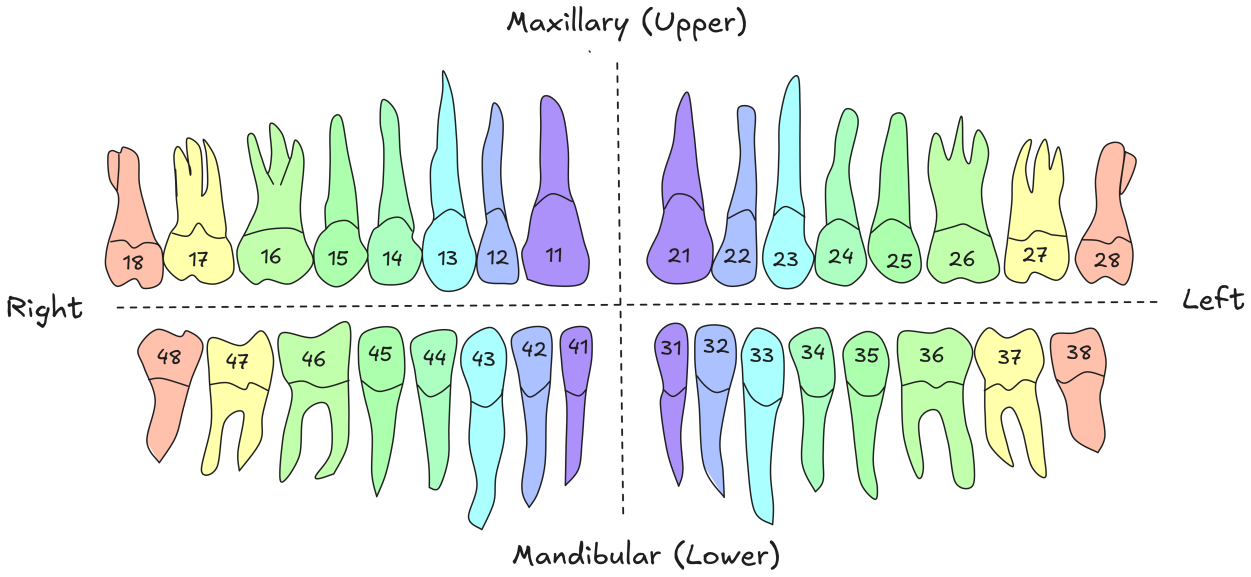
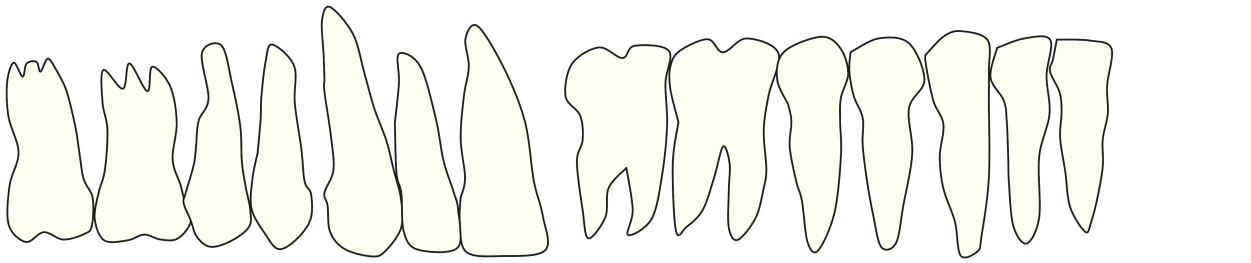


Figure 2.4: FDI tooth numbering (indexing) system. Tooth positions are defined with respect to the patient's anatomical left and right sides, rather than the observer's perspective.



17,27	16,26	15,25	14,24	13,23	12,22	11,21	47,37	46,36	45,35	44,34	43,33	42,32	41,31	
2.0	4.0	5.0	5.5	6.0	5.5	6.0	3.5	3.5	4.5	5.0	5.5	5.0	5.0	+ 1.5 mm
2.0	3.5	4.5	5.0	5.5	5.0	5.5	3.0	3.0	4.0	4.5	5.0	4.5	4.5	+ 0.5 mm
2.0	**3.0**	**4.0**	**4.5**	**5.0**	**4.5**	**5.0**	**2.5**	**2.5**	**3.5**	**4.0**	**4.5**	**4.0**	**4.0**	**mean**
2.0	2.5	3.5	4.0	4.5	4.0	4.5	2.0	2.0	3.0	3.5	4.0	3.5	3.5	- 0.5 mm
2.0	2.0	3.0	3.5	4.0	3.5	4.0	2.0	2.0	2.5	3.0	3.5	3.0	3.0	- 1.0 mm

Figure 2.5: The diagram illustrates the recommended vertical distances (in millimeters) for orthodontic bracket placement, measured from the incisal edge or occlusal cusp to the bracket center. The chart provides the mean heights for the maxillary and mandibular teeth, ranging from 2.5 mm to 5.0 mm, alongside incremental adjustments (+1.5 mm to -1.0 mm) used to customize the installation based on patient needs.

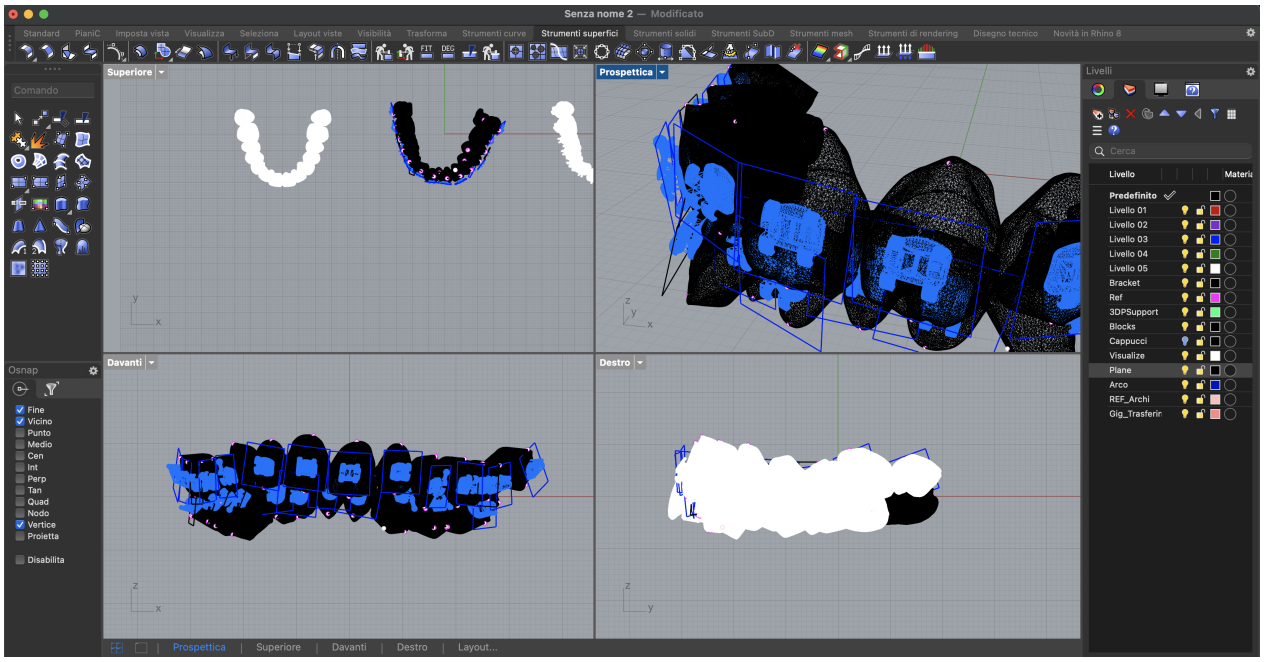


Figure 2.6: 3D modelling environment used by orthodontic specialists to replicate the traditional manual procedure of vertical bracket placement.

2.5 The STL File Format

From a technical perspective, intra-oral scans must be stored in a digital format capable of accurately representing complex three-dimensional surface geometry while remaining compatible with downstream processing pipelines, including segmentation, geometric analysis, and machine learning models. Among the various available 3D formats, the STL (Stereolithography) file format is the most widely adopted standard in orthodontics and dental CAD/CAM systems, largely due to its simplicity, broad software support, and scanner compatibility.

An STL file represents a 3D object as a polygonal surface mesh composed exclusively of triangular facets. Each triangle is defined by three vertices in three-dimensional Euclidean space and an associated outward-pointing surface normal vector. Formally, a single triangular facet can be described as:

$$T = \boldsymbol{n}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \tag{2.1}$$

where $\boldsymbol{v}_i \in \mathbb{R}^3$ denote the Cartesian coordinates of the triangle vertices and $\boldsymbol{n} \in \mathbb{R}^3$ represents the unit normal vector, typically computed as the normalized cross product of two triangle edges. The

normal encodes the local surface orientation and is primarily used for visualization and rendering purposes rather than geometric reconstruction.

The complete STL mesh is obtained by aggregating a large set of such triangular facets, which together approximate the continuous surface of the scanned dental anatomy. Increasing the number of triangles improves surface fidelity but also increases file size and computational cost. Modern intra-oral scanners generate high-resolution meshes to capture fine anatomical details such as cusps, fissures, and interproximal surfaces, which are critical for orthodontic planning.

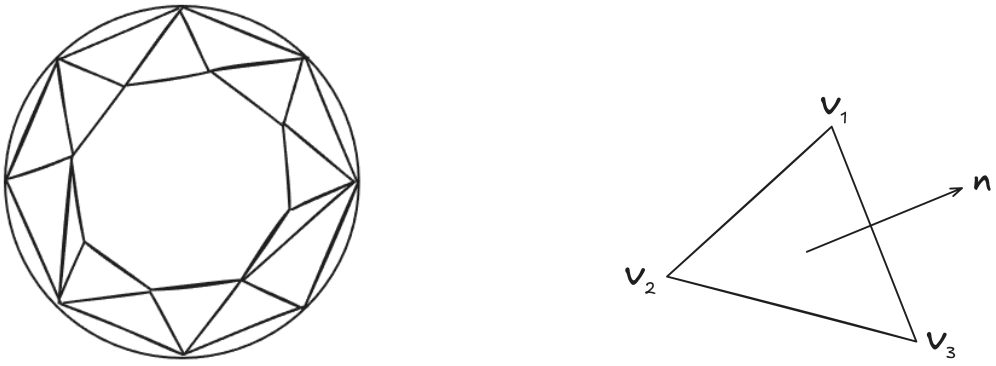
A notable characteristic of the STL format is its lack of explicit topological and semantic information. Although triangles may share vertices and edges implicitly, adjacency relationships are not explicitly encoded. Furthermore, the format does not provide any metadata or labels indicating anatomical structures, such as individual teeth, gingiva, or occlusal surfaces. As a consequence, STL files represent purely geometric information, and any higher-level understanding of the scan, such as tooth segmentation or landmark identification, must be inferred through additional processing steps.

In clinical practice, intra-oral scans are typically stored as binary STL files, which offer reduced file size and faster parsing compared to ASCII STL representations. These files describe a surface mesh that can be visually interpreted as a wireframe structure composed of interconnected triangular facets. As illustrated in Fig. 2.7, the left panel shows a wireframe representation of a scanned object, while the right panel presents a zoomed-in view of a single triangular facet belonging to that mesh. Due to the high resolution of modern scanners, a single STL file may contain several hundred thousand to millions of triangles. While this level of detail is advantageous for accurate orthodontic analysis, it also introduces challenges related to memory usage, noise sensitivity, and computational complexity.

Relation to point clouds. Although STL files are surface-based representations, many geometric processing and machine learning pipelines operate on point clouds. A point cloud is an unordered set of points in three-dimensional space. Unlike meshes, point clouds do not encode explicit connectivity, surface normals, or faces, and therefore constitute a more minimal geometric representation.

Point clouds can be directly acquired from 3D scanners or derived from STL meshes by sampling vertices or uniformly resampling points over the triangular surface. This conversion is particularly

relevant for deep learning approaches that are specifically designed to operate on unordered point sets. While point clouds discard explicit surface connectivity, they offer advantages in terms of simplicity, robustness to topological inconsistencies, and compatibility with modern neural network architectures.



(a) Wireframe (b) Triangle

Figure 2.7: Example of a wireframe representation (a) and a zoomed-in view of a single triangle, highlighting its vertices and normal vector (b).

3. Related Work

3.1 Introduction

Recent advances in three-dimensional (3D) acquisition technologies [31, 1] have played a key role in enabling the automation of orthodontic workflows. Modern intra-oral scanners generate high-resolution digital representations of dental anatomy, typically in the form of polygonal meshes or point clouds. These representations allow for the direct application of automated algorithms and learning-based methods to tasks such as tooth segmentation, landmark detection, and geometric analysis.

This chapter reviews the relevant literature by progressively introducing the most common learning tasks performed on 3D data, together with modern deep learning architectures designed for three-dimensional representations. Particular emphasis is placed on point-based and transformer-based models, which are especially well suited to the processing of intra-oral scans. This structured overview provides the conceptual background necessary to position the proposed approach within the broader context of 3D medical image analysis and dental computing.

3.2 Learning Tasks on 3D Meshes and Point Clouds

Three-dimensional geometric data are commonly represented as polygonal meshes, voxel grids, multi-view projections, or point clouds. Among these, meshes and point clouds are particularly prevalent in medical and dental applications due to their ability to accurately encode surface geometry. Regardless of the chosen representation, a number of core learning tasks recur across 3D computer vision, including:

- **Classification and Shape Analysis**: Shape classification aims to assign a single semantic label to an entire 3D object, such as identifying an anatomical structure or categorizing an object class. Early works [40, 29] in 3D deep learning often focused on this task as a benchmark for evaluating global shape representations. Closely related tasks include

shape retrieval and similarity analysis, where embeddings are learned to reflect geometric or anatomical similarity [5] between objects.

- **Semantic Segmentation**: Semantic segmentation is one of the most fundamental tasks in 3D geometric learning. It consists in assigning a semantic label to each primitive element of the representation, such as a vertex in a mesh, a voxel in a grid, or a point in a point cloud. In medical contexts, this typically corresponds to identifying anatomical regions or tissue types. A prominent example is the work by Bolelli et al [4], which establishes a benchmark for state-of-the-art architectures in dental imaging. Their study evaluates transformer-based models like TransU-Net [6] and UNETR++ [33] against traditional convolutional and hybrid Mamba-based [12] frameworks for the voxel-level segmentation of 42 distinct anatomical classes. By leveraging the long-range dependency modeling inherent in attention-based and state-space architectures, this work addresses the significant anatomical variability and low contrast typical of Cone-beam computed tomography (CBCT) scans.

- **Keypoint and Landmark Detection** Another important class of tasks concerns the detection of sparse but semantically meaningful points, commonly referred to as keypoints or landmarks. In dental applications, landmarks correspond to clinically relevant locations on tooth crowns, such as cusp tips or bracket placement references. These tasks require models to capture fine-grained local geometry while maintaining awareness of global anatomical context. Further details on this task are presented in Section 3.5.

3.3 Deep Learning Architectures for 3D Data

3.3.1 Introduction

Deep learning architectures for three-dimensional data are rooted in developments originally introduced in the two-dimensional image domain. In 2D computer vision, convolutional neural networks (CNNs) have long served as the dominant paradigm for tasks such as image classification, object detection, and semantic segmentation. Architectures such as Fully Convolutional Networks (FCNs) enabled dense, per-pixel prediction by replacing fully connected layers with convolutional ones,

while encoder–decoder designs with skip connections, exemplified by U-Net [32], proved particularly effective in preserving spatial detail for biomedical image segmentation. Subsequent models, including DeepLab [7], further enhanced performance by enlarging the receptive field through atrous convolutions.

More recently, attention-based models and vision transformers (ViTs) [9] have challenged the dominance of purely convolutional architectures in 2D vision. By modeling global interactions between image patches via self-attention, transformer-based approaches enable the capture of long-range dependencies that are difficult to represent with local convolutional filters alone. Hybrid architectures and fully transformer-based models, such as SegFormer [41], have demonstrated strong performance in dense prediction tasks while offering improved scalability and architectural flexibility.

Extending these ideas from 2D images to 3D geometric data, however, is non-trivial. Unlike images, three-dimensional representations such as point clouds and meshes are unordered, irregular, and lack a canonical grid structure. These properties prevent the direct application of standard convolutional operators and require architectures that are invariant to point ordering while remaining sensitive to spatial relationships. As a result, the development of deep learning models for 3D data has followed distinct architectural paths, including voxel-based, multi-view, and point-based approaches.

Among these, point-based neural networks have emerged as a particularly effective solution for processing raw geometric data, as they operate directly on point sets without intermediate discretization. More recently, transformer-based architectures have been adapted to the point cloud domain, leveraging self-attention mechanisms to model both local geometry and global context. The following sections review these architectures, with a focus on point-based networks and attention-driven models that are especially relevant for the analysis of intra-oral scans.

3.3.2 Point-Based Neural Networks

Point clouds represent 3D geometry as unordered sets of points sampled from object surfaces. Their simplicity and efficiency make them particularly attractive for processing intra-oral scans, which are often stored as STL meshes that can be readily sampled into point sets.

PointNet [29] was the first architecture to process point clouds directly without intermediate

voxelization or projection. It applies shared multilayer perceptrons (MLPs) to individual points and aggregates global information using symmetric functions to ensure permutation invariance. While effective for global tasks, PointNet lacks explicit modeling of local geometric structure.

PointNet++ [30] addresses this limitation by introducing hierarchical feature learning through neighborhood grouping and local aggregation. By recursively capturing local context at multiple scales, PointNet++ significantly improves performance on segmentation tasks. However, both architectures rely on manually defined neighborhood operations and have limited ability to model long-range dependencies.

Building upon hierarchical feature extraction, recent work has explored self-supervised point cloud upsampling to handle sparse and irregular point sets without requiring dense ground-truth supervision [21]. In particular, SPU-Net introduces a coarse-to-fine reconstruction framework that first downsamples input patches and then upsamples them hierarchically to generate dense and uniform point clouds. The method combines graph convolutional networks with self-attention to capture both local and inter-region context, and uses a hierarchical learnable folding strategy for feature expansion. A self-projection optimization further refines the generated points to lie on the underlying surface, enabling robust reconstruction from sparse inputs. Such approaches are especially relevant for intra-oral scans, which often exhibit irregular point distributions and incomplete sampling.

3.3.3 Attention-Based and Transformer Architectures

Motivated by the success of self-attention in natural language processing [34] and 2D vision [9], transformer-based architectures have been adapted to 3D point cloud processing. In contrast to fixed local aggregation schemes, self-attention allows each point to dynamically attend to other points based on both spatial relationships and learned feature similarity.

Early point cloud transformers restricted attention to local neighborhoods to control computational cost. A major milestone was achieved with the Point Transformer [43], which introduced vector attention mechanisms explicitly designed for 3D geometry. By incorporating relative positional encodings derived from point coordinates, Point Transformer models effectively captures both local and global geometric context.

Recent developments have focused on improving scalability and efficiency. Among these, Point

Transformer V3 [39] represents a significant advancement by rethinking how attention is applied to large-scale point clouds. The following section examines this architecture in greater detail, as it plays a fundamental role in the proposed approach.

3.3.4 Point Transformer V3

Point Transformer V3 (PTv3 [39]) is a recent advancement in transformer-based architectures for point cloud processing, designed with the explicit goal of overcoming the long-standing trade-off between accuracy and computational efficiency in 3D perception. Unlike earlier point transformers, which focus on increasingly complex attention mechanisms to improve local geometric modeling, PTv3 adopts a fundamentally different design philosophy: it prioritizes simplicity and efficiency in order to enable large-scale receptive fields and effective model scaling.

The proposed architecture introduces three core design elements: point cloud serialization through space-filling curves, serialized patch-based self-attention, and sparse convolution-based positional encoding. Each of these components is described in detail in the remainder of this section. An overview of the complete architecture is shown as a reference in Figure 3.1.

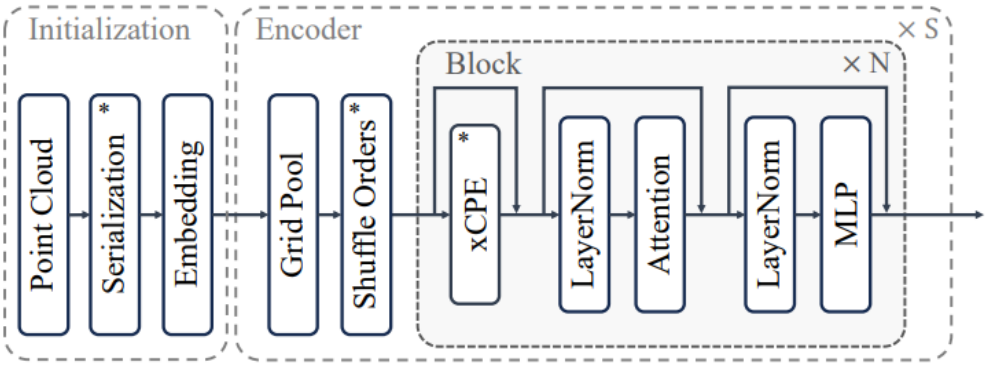


Figure 3.1: Point Transformer V3 [39] architecture.

Point cloud serialization via space-filling curves Instead of treating a point cloud as an unordered set, PTv3 converts it into a structured one-dimensional sequence using space-filling curves [24], thereby enabling the direct application of self-attention, which is inherently defined over one-dimensional sequences of tokens. Space-filling curves, such as Z-order and Hilbert curves, map multi-dimensional spatial coordinates to a one-dimensional ordering while preserving spatial prox-

imity to a reasonable extent.

Formally, each point is assigned a serialization code derived from its discretized spatial coordinates. Sorting points according to this code yields a sequence in which neighboring points in the sequence are likely to be spatially close in 3D space. By imposing a deterministic one-dimensional ordering on the point cloud, serialization allows self-attention to operate on points in a manner analogous to tokens in natural language or image patches in vision transformers.

PTv3 employs multiple serialization patterns, including standard and axis-transposed variants of Z-order and Hilbert curves, to capture diverse spatial relationships. Across successive attention layers, these serialization patterns are alternated or randomly shuffled to mitigate biases introduced by any single ordering. This serialization step removes the need for explicit neighborhood queries such as K-nearest neighbors (KNN), which constitute a major computational bottleneck in earlier point transformer architectures.

The three-dimensional Hilbert curve, illustrated in Figure 3.2, demonstrates a continuous, space-filling trajectory that maps one-dimensional indices to a higher-dimensional volume while maintaining high spatial locality.

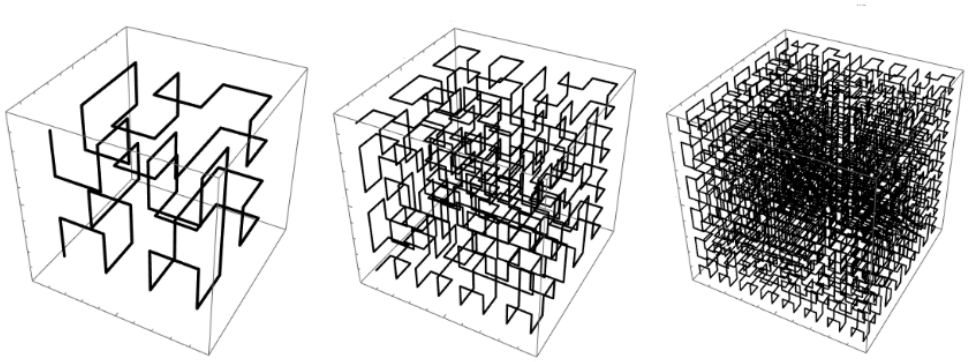


Figure 3.2: Progressively denser space filling Hilbert curve.

Serialized patch-based self-attention The core of the PTv3 architecture is a serialized patch-based attention mechanism that leverages the structured 1D order derived from space-filling curves. Unlike its predecessor, PTv2 [38], which utilized grouped vector attention and K-nearest neighbor (KNN) queries, PTv3 adopts a more efficient strategy. Points in the serialized sequence are grouped into non-overlapping patches of a fixed size, and self-attention is performed exclusively within each

individual patch.

This design choice facilitates a significant expansion of the model's receptive field. While traditional point transformers are often limited to local neighborhoods of approximately 16 points, PTv3 can efficiently scale its patch size to 1024 points or more. To ensure effective information propagation across the entire point cloud and to mitigate boundary effects introduced by non-overlapping patches, PTv3 adopts dynamic patch interaction strategies. In particular, the serialization order of points is not kept fixed across attention blocks. Instead, the specific ordering pattern (e.g., Z-order or Hilbert curve) is cyclically varied between successive layers, allowing the model to capture complementary spatial contexts at different depths. Furthermore, the sequence of serialization patterns is randomly shuffled before being processed by the attention layers, which prevents the network from overfitting to a single spatial traversal and improves its generalization capability.

Positional encoding via sparse convolution In Point Transformer V3, positional information is injected into point features using sparse convolution [20] rather than explicit relative positional encoding. Sparse convolution extends standard convolution to sparse voxel grids by applying convolutional kernels only at occupied spatial locations, making it well suited for irregular 3D point clouds.

To this end, the point cloud is discretized into a sparse voxel grid [20] by quantizing point coordinates, and points falling within the same voxel are aggregated to form voxel-level features. Sparse convolution is applied to these features, enabling local geometric context to be captured through interactions between neighboring voxels. The resulting voxel features are then mapped back to individual points, yielding one spatially conditioned feature vector per point.

This process provides a learnable positional encoding, referred to as extended conditional positional encoding (xCPE), which encodes absolute spatial location and local geometry. By introducing positional information prior to self-attention, sparse convolution removes the need for computationally expensive pairwise relative positional encoding, allowing attention to scale efficiently to large serialized point sequences.

3.4 Segmentation of Intra-Oral Scans

Semantic segmentation of intra-oral scans involves assigning a label to each element of a digital dental model, typically with the goal of identifying individual teeth and surrounding anatomical structures. This task constitutes a foundational step in digital orthodontics, as errors in segmentation propagate to downstream processes such as landmark localization and bracket placement. The challenges of intra-oral scan segmentation include complex tooth morphology, close inter-tooth contact, and variations in scan quality. Modern approaches increasingly rely on point-based and transformer-based architectures, which can directly operate on raw geometric data while capturing both local surface details and global dental arch structure. As a result, transformer-based point cloud models have emerged as a powerful tool for accurate and robust dental segmentation.

Several learning-based approaches have been proposed for the semantic segmentation of intra-oral scans, aiming to separate teeth and surrounding anatomical structures from three-dimensional dental models. In [35], the authors formulate the problem using a convolutional neural network operating on transformed representations of the dental geometry, demonstrating that learned features can outperform purely geometry-based heuristics in challenging anatomical regions. In [42], the authors propose a point-based segmentation framework that operates directly on sampled surface points from intra-oral scans. By leveraging local neighborhood aggregation and hierarchical feature learning, the method achieves improved boundary delineation between adjacent teeth. However, the approach is evaluated on a proprietary dataset and does not provide publicly available code or pretrained weights, limiting its reproducibility and practical reuse.

A different strategy is explored in [44], where dental meshes are modeled using graph-based representations and processed with graph convolutional networks. This formulation enables explicit modeling of surface connectivity and local geometric relationships, resulting in accurate segmentation in regions with subtle geometric transitions. Despite these promising results, the lack of released implementation details and trained models again restricts the applicability of the method beyond the original study.

Lian et al. proposed MeshSegNet [18], an end-to-end framework that automates the labeling of raw 3D dental surfaces by hierarchically extracting multi-scale local geometric features through graph-constrained learning modules and a dense fusion strategy. By utilizing a comprehensive 15-

D input vector (which includes vertex coordinates, cell normals, and relative positions) the model effectively captures fine-grained local context and achieves high-precision segmentation even in challenging regions like incompletely scanned molars.

Overall, while many works in this field demonstrate the feasibility of deep learning for intra-oral scan segmentation, they often lack the transparency needed for clinical adoption due to restricted access to code. In contrast, MeshSegNet distinguishes itself by providing an open-source implementation and trained models, establishing it as a reliable and reproducible baseline for the community. This approach to open research is also followed by ToothInstanceNet [26], a more recent architecture that similarly provides public code and weights. Due to its specific relevance to dental instance-level understanding and its accessibility for deployment, a more in-depth description of ToothInstanceNet is provided in Section 3.5 and 3.6.

3.5 The 3D-Teethland Challenge

The 3DTeethLand Challenge [23] was organized to advance research on automated understanding of intra-oral scans by focusing on tooth landmark detection and classification. While previous benchmarks primarily addressed tooth instance segmentation and labeling, tasks like orthodontic planning require more detailed geometric information, notably the localization of anatomical tooth landmarks. The challenge dataset builds upon the publicly available Teeth3DS collections and provides sparse landmark annotations for a subset of scans, thereby enabling supervised learning approaches for this task.

The objective of the challenge is to detect and classify six types of landmarks on 3D intra-oral scans: mesial, distal, facial, inner, outer, and cusp landmarks. Predictions are evaluated using distance-based metrics over multiple tolerance thresholds. Among the submitted methods, ToothInstanceNet [26] achieved first place on the final test leaderboard, demonstrating strong performance across all landmark categories. Its success is primarily attributed to a carefully designed two-stage architecture that separates global contextual reasoning from local high-resolution geometric analysis.

3.6 ToothInstanceNet

ToothInstanceNet is the method that achieved first place in the 3DTeethLand Challenge (2024) and is designed to predict comprehensive tooth-level information from intra-oral scans, including tooth instance segmentation, anatomical tooth labeling, and landmark detection. The model follows a two-stage architecture that explicitly separates global, context-aware reasoning from local, high-resolution geometric analysis.

The first stage operates on a downsampled representation of the full intra-oral scan and performs tooth instance segmentation and labeling. Each point is embedded in a spatial representation by predicting a seed score, an offset vector toward the center of the corresponding tooth, and a bandwidth encoding the spatial extent of the instance. Tooth instances are recovered through an iterative clustering procedure that models each tooth as a three-dimensional Gaussian distribution. Tooth labeling is performed at the instance level by aggregating point-wise features using masked average pooling, followed by a multi-layer perceptron that predicts the tooth class. This stage is trained end-to-end using a composite loss function that combines a spatial embedding loss for instance separation, cross-entropy and focal losses for tooth classification, and a feature consistency regularization that encourages coherence of point features within each tooth instance. The overall loss for the first stage can be written as

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{SE}} + \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Focal}} + \mathcal{L}_{\text{consistency}}.$$

The second stage focuses on anatomical landmark detection and operates on high-resolution point cloud crops extracted around each predicted tooth instance. For each landmark class, the network predicts per-point distances to the closest landmark and offset vectors pointing toward the landmark location, thereby generating dense landmark proposals across the tooth surface. These predictions are supervised using a combination of distance regression, offset regression, and separation losses, together with a binary segmentation loss that refines the tooth mask at high resolution. The final landmark positions are obtained during inference by clustering the landmark proposals using a weighted density-based algorithm. The loss of the second stage is defined as

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{seg}} + \sum_{k=1}^{K} \mathcal{L}_{\text{landmark}}^{(k)}.$$

Although each stage of ToothInstanceNet is trained end-to-end, the full pipeline is not optimized jointly across stages. Instead, the model adopts a modular training strategy in which the output of the first stage is used to generate inputs for the second stage. This choice improves training stability and reflects a hybrid learning paradigm that combines deep neural networks with geometric clustering and rule-based post-processing. Such a design is particularly well suited for dental applications, where strong anatomical priors can be exploited to ensure geometrically plausible and clinically meaningful predictions.

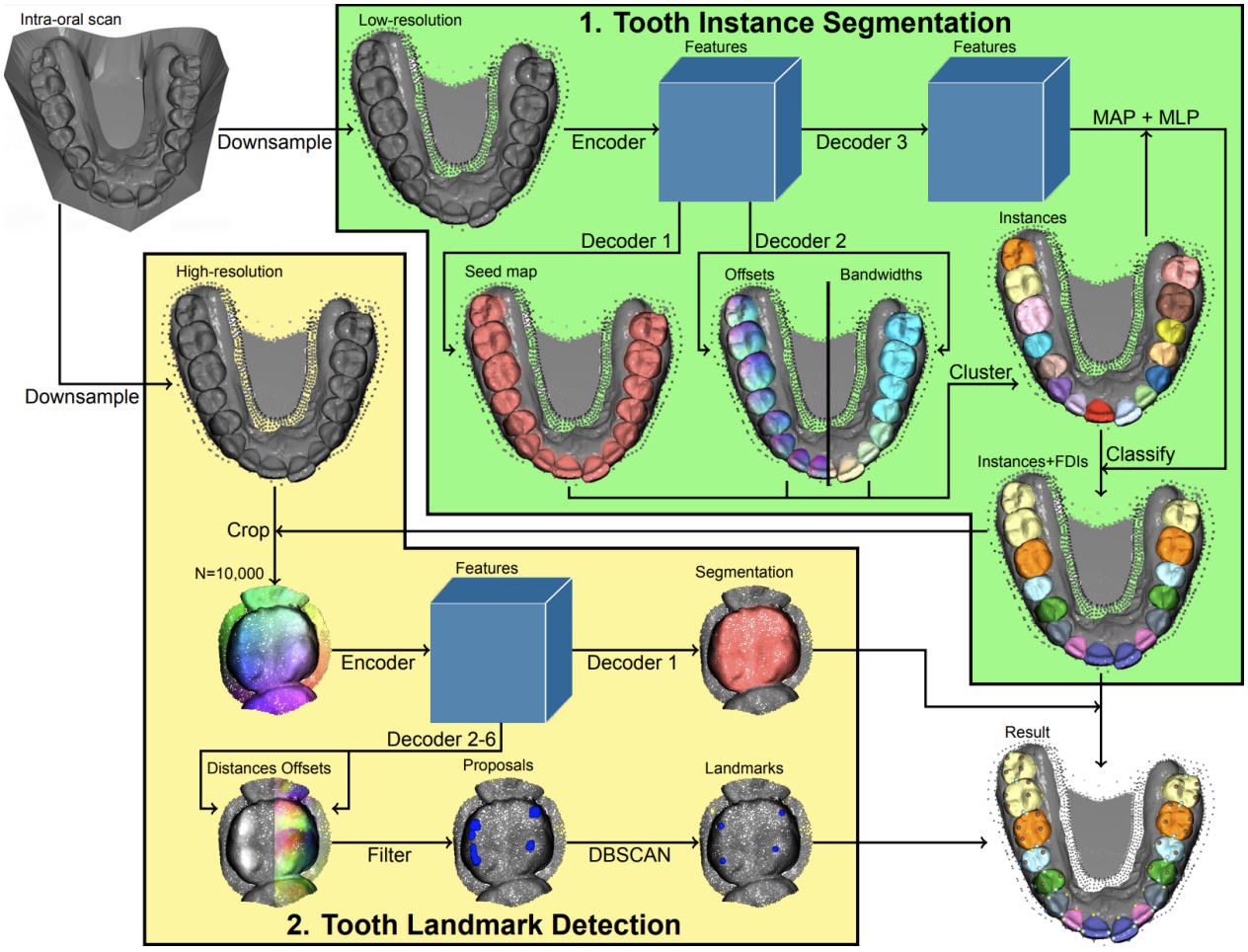


Figure 3.3: Overview of the ToothInstanceNet [26] architecture. The pipeline consists of a two-stage framework in which a low-resolution, large-context network performs tooth instance segmentation and labeling on the full intra-oral scan, followed by a high-resolution, tooth-centric network that refines individual tooth segmentations and predicts anatomical landmarks using distance–offset regression and clustering.

4. Datasets

4.1 Introduction

This chapter describes the datasets curated for the development and evaluation of the proposed orthodontic pipeline. The system addresses two tightly coupled tasks, namely the semantic segmentation of full intra-oral scans and the prediction of tooth-level landmarks for automatic bracket placement. To support these objectives, three complementary datasets were collected and processed, each targeting a specific stage of the pipeline.

The first dataset is a high-quality, expert-annotated collection specifically designed to supervise the prediction of clinically relevant landmarks and bracket positioning at the tooth level. The second dataset is a large-scale corpus of full intra-oral scans assembled to train a custom semantic segmentation model, which is subsequently employed as a preprocessing component of the overall system in production.

In addition to these datasets, a third dataset, hereafter referred to as the *Melted dataset*, was constructed to bridge full-arch representations and isolated tooth analysis. This dataset is obtained by using the ToothInstanceNet model to segment intra-oral scans in order to separate teeth from gingival tissue and to identify individual tooth instances. Each tooth, labeled according to its FDI index, is extracted, centered at its center of mass, and isotropically scaled to fit within a unit sphere, while preserving the original orientation inherited from the source scan. This representation enables tooth-centric learning while maintaining a consistent spatial relationship with the original intra-oral geometry.

For all datasets, particular attention was devoted to geometric consistency, annotation reliability, and the traceability of all applied transformations, ensuring reproducibility and robustness across training and evaluation stages.

4.2 The Brackets Dataset

4.2.1 Dataset Acquisition

The bracket placement dataset was collected in collaboration with an orthodontic specialist studio in Ferrara (Italy). It comprises a total of 312 intra-oral scans corresponding to 158 patients. For each patient, scans of both the upper (maxillary) and lower (mandibular) arches were acquired whenever available; however, in a limited number of cases, only one of the two arches was present.

Each scan is stored in STL format and represents a high-resolution digital impression of the patient's dentition, acquired with an average sampling density of approximately 100,000 vertices. The dataset includes scans acquired under real clinical conditions and therefore exhibits natural variability in terms of anatomy, acquisition noise, and scan completeness.

In addition to the raw geometric data, each patient is associated with a metadata file stored in JSON format. This file contains orthodontic annotations provided by expert clinicians and encodes all information required for supervised learning of bracket placement.

4.2.2 Annotation Structure

For each tooth requiring bonding, the associated metadata includes the following elements:

- **ptTop**: a 3D point corresponding to one extremity of the bracket reference annotation;
- **ptBottom**: a second 3D point defining the complementary extremity of the annotation;
- **Base Plane**: a local reference plane corresponding to the plane on which the 3D mesh of the bracket lies. The base plane is defined by:
  - an **origin** point located on the bracket base,
  - an **X-axis** direction vector lying within the plane,
  - a **Y-axis** direction vector lying within the plane and orthogonal to the X-axis.

Together, these elements define both the position and the local orientation of the desired bracket placement for each annotated tooth, thus encoding clinically meaningful information beyond a single landmark.

Furthermore, each metadata file contains a field named **scanTransformMatrix**, which represents a rigid Euclidean transformation composed of a rotation and a translation. This matrix maps the raw STL scan into a common reference frame, ensuring that the geometric data and the annotated points are expressed in the same coordinate system and that maxillary and mandibular scans are consistently registered. Since this coordinate system does not align with the one adopted for this study, a subsequent preprocessing phase is required. The specific geometric transformations and normalization procedures involved in this stage are detailed in Section 4.2.3.

4.2.3 Normalization and Alignment

To make the collected scans suitable for large-scale automated processing and neural network training, a sequence of geometric normalization procedures is applied. These procedures standardize the coordinate ranges and spatial representations of the data to improve numerical stability and to conform to common practices in deep learning pipelines, while preserving the exact spatial alignment between each scan and its associated annotations.

Centering Each scan was translated such that its center of gravity coincides with the origin of the reference coordinate system. The center of gravity was computed from the mesh vertices, and the same translation was applied to all annotated points contained in the metadata files, including *ptTop*, *ptBottom*, and the base plane parameters.

Scale normalization Subsequently, each scan was isotropically scaled to fit within a unit sphere of radius 1 centered at the origin. This normalization mitigates differences in absolute scan dimensions across patients and acquisition devices. The scaling factor was consistently applied to both the scan geometry and all annotation points.

Orientation standardization An additional rigid rotation was applied to enforce a standardized orientation, following the convention adopted in previous work. After this transformation:

- the **X axis** points towards the patient's right side;
- the **Y axis** points outward from the oral cavity;

- the **Z axis** points towards the patient's skull;

This reference frame configuration is hereafter referred to as the **standard orientation**.

Annotation refinement The original distinction between *ptTop* and *ptBottom* was determined by the temporal order of manual clicks during annotation and was therefore not always anatomically consistent. To enforce a semantically meaningful labeling, a post-processing relabeling strategy based on geometric criteria was adopted.

Specifically, the relative position of the two points along the standardized Z axis was used to infer their anatomical role. For upper arches, the point with the lower Z coordinate was classified as *incisal*, while the point with the higher Z coordinate was assigned as *gingival*; for lower arches, this criterion was inverted. This procedure ensured anatomically coherent and consistent point annotations across the entire dataset.

The result of this normalization and annotation pipeline is illustrated in Figure 4.1. The figure showcases two registered scans (maxillary and mandibular) centered at their collective center of mass and scaled within a unit sphere. The visualization highlights, from two different points of view, the standardized orientation of the arches alongside the refined orthodontic annotations, specifically the bracket placement planes and the semantically labeled incisal and gingival points.

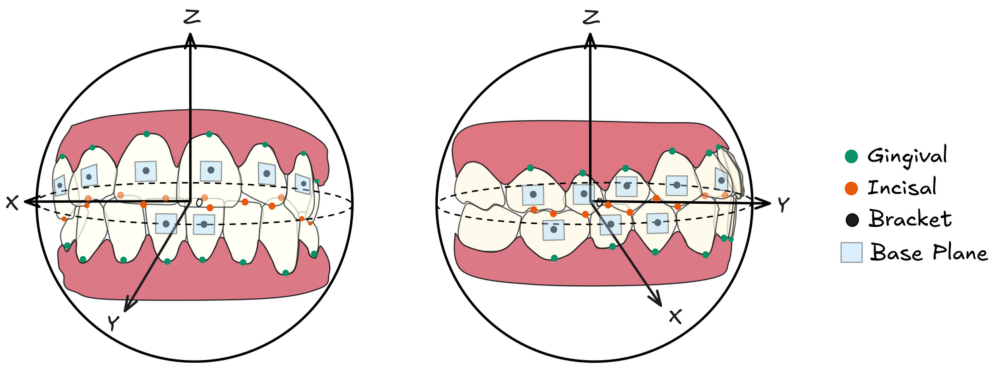


Figure 4.1: Illustration of a normalized dataset sample from the Brackets dataset. The maxillary and mandibular scans are registered and scaled to fit within a unit sphere centered at the origin. For each tooth, the visualization shows the bracket placement base planes and the corresponding *ptTop* and *ptBottom* landmarks, now consistently relabeled as *incisal* and *gingival* points according to the standardized Z-axis orientation.

4.3 Semantic Segmentation Dataset

4.3.1 Motivation and Scope

In addition to precise landmark supervision, the proposed pipeline requires reliable tooth-level segmentation of full intra-oral scans. Since segmentation errors directly affect downstream landmark prediction and bracket placement, a dedicated large-scale dataset was assembled to train a robust semantic segmentation model suitable for production use.

Manual annotation of tooth-level segmentation at this scale would be prohibitively time-consuming and clinically expensive. Therefore, a weakly supervised strategy based on high-quality pseudo-labels was adopted.

4.3.2 Dataset Composition

The segmentation dataset was constructed by aggregating multiple proprietary and publicly available sources, resulting in a large and heterogeneous collection of intra-oral scans. The final dataset includes 3,590 scans, comprising the Teeth3DS dataset, scans previously collected from the Ferrara orthodontic studio for earlier studies, a specific subset of the scans described in the previous chapter, and a publicly available dataset consisting of intra-oral scans and corresponding CBCT volumes [17]. By combining these diverse data sources, the final collection encompasses a wide spectrum of dental morphologies and clinical conditions, providing a robust foundation for the training and evaluation of the proposed segmentation and placement models.

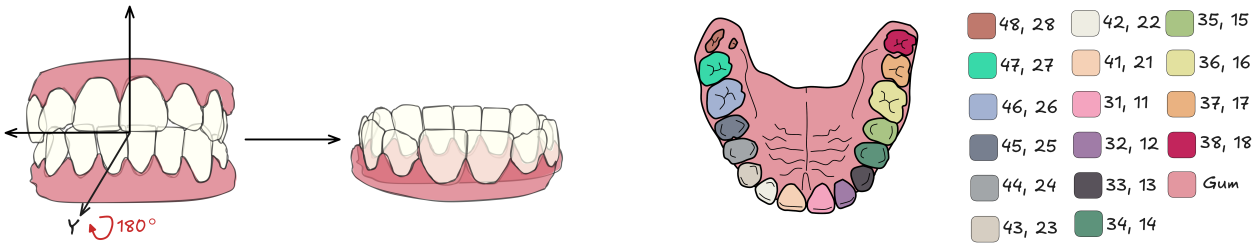
All scans in the segmentation dataset were automatically annotated using ToothInstanceNet (see section 3.6). Although computationally expensive and unsuitable for real-time deployment, ToothInstanceNet provides highly accurate tooth instance segmentations and therefore represents a reliable source of pseudo–ground truth.

The generated instance labels were converted into per-vertex semantic labels, distinguishing gingiva and individual teeth. These labels were then used to supervise the training of a lightweight semantic segmentation network, enabling a significant reduction in inference time while maintaining competitive accuracy.

4.3.3 Class Definition

The segmentation task is formulated as a multi-class classification problem with 17 classes. Class 0 corresponds to gingival tissue, while classes 1 to 16 correspond to individual tooth instances along half of a dental arch.

To reduce label complexity, we apply an extra 180° rotation around the Y axis with respect to the standard orientation to the maxillary(upper) scan, such that anatomically corresponding teeth in upper and lower arches are spatially aligned. As a result, the segmentation model is agnostic to the arch type, and full FDI indices are recovered a posteriori based on the known scan metadata. Figure 4.2 illustrates the input and the output of the segmentation model.



(a) Before segmentation

(b) After segmentation

Figure 4.2: Preprocessing and segmentation of intra-oral scans. (a) Spatial alignment process where the maxillary (top) arch is rotated 180° around the Y-axis to achieve anatomical overlap with the mandibular (bottom) arch. (b) The resulting multi-class segmentation mask; class 0 is reserved for the gingiva (gum), while classes 1 through 16 represent specific tooth instances. The legend denotes the FDI index for each instance, where class 1 corresponds to teeth 48 and 28, and class 16 corresponds to teeth 38 and 18.

4.4 The Melted Dataset

4.4.1 Motivation and Scope

The Melted dataset was specifically designed to enable efficient and high-precision training of the proposed bracket placement model on isolated tooth geometries. In clinical practice, orthodontists position brackets by focusing on one tooth at a time, performing fine-grained adjustments based on

tooth morphology and orientation, with only minor refinements depending on the specific treatment plan. With the exception of such treatment-specific adjustments, which are outside the scope of this thesis, the goal of the proposed model is to replicate this tooth-level precision.

By training directly on single tooth meshes, the model is encouraged to implicitly recognize the tooth type and to infer the correct placement of orthodontic landmarks purely from geometric cues. This formulation avoids the complexity of full-arch representations and allows the learning process to focus on local morphology, which is the primary source of information used during manual bracket positioning.

An additional motivation for this dataset is computational efficiency. Since all tooth instances are pre-extracted, the bracket prediction model can be trained without performing semantic or instance segmentation during training, resulting in a simpler and faster training pipeline.

4.4.2 Dataset Construction

The Melted dataset is constructed by processing the scans belonging to the Brackets dataset described in Section 4.2. Each intra-oral scan is segmented using ToothInstanceNet, the instance segmentation framework introduced in the previous chapter. This model separates dental crowns from gingival tissue and assigns each tooth a unique instance label according to the FDI notation system.

Following segmentation, each tooth instance is extracted from the full-arch mesh as an independent geometric entity. To ensure geometric consistency, only the largest connected component of each extracted tooth mesh is retained, removing small disconnected fragments that may arise from segmentation artifacts. In addition, a limited number of tooth instances affected by severe segmentation errors were manually discarded. This manual filtering step was introduced to maintain a high-quality collection of tooth meshes and to avoid introducing corrupted samples that could negatively affect training stability and performance. As shown in Figure 4.3, the resulting distribution of samples across the different tooth classes is quite uniform, ensuring that the model is exposed to a balanced variety of dental morphologies during the learning phase.

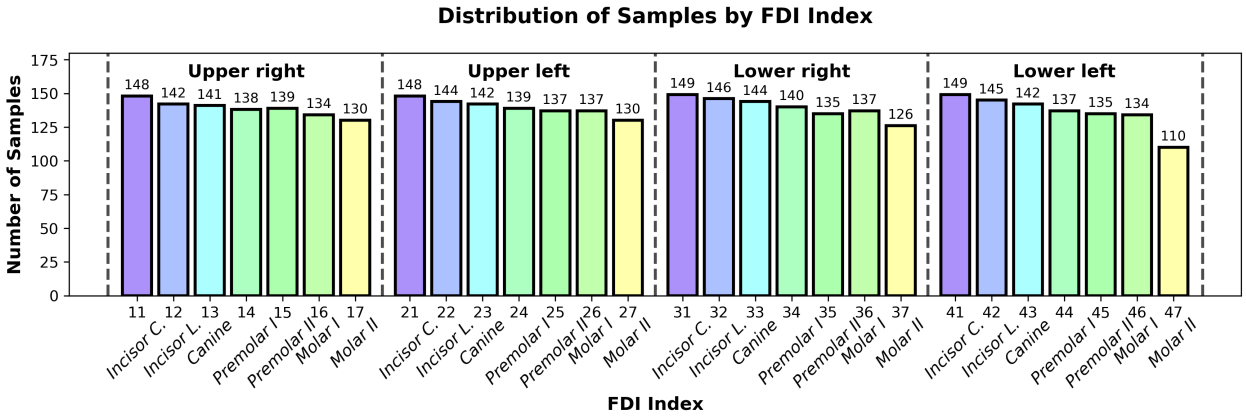


Figure 4.3: Distribution of tooth samples across 28 FDI tooth classes in the Melted dataset. Bars are colored by tooth type (central incisor through second molar), matching the color scheme in Figure 2.4. Dashed lines separate the four oral quadrants.

4.4.3 Geometric Normalization

To standardize the input representation while preserving clinically relevant information, a sequence of geometric normalization steps is applied to each extracted tooth. These operations are designed to reduce unnecessary variability while maintaining a consistent and traceable relationship between geometry and annotations.

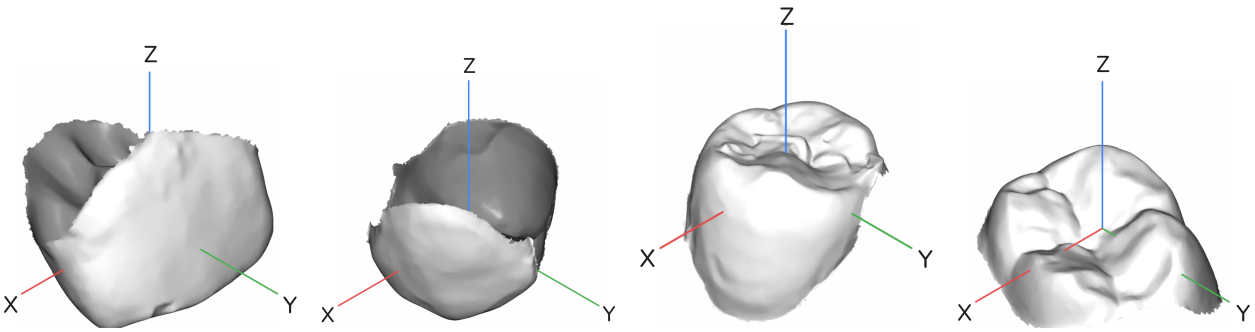
Centering Each tooth mesh is translated such that its center of mass coincides with the origin of the coordinate system.

Scale normalization The mesh is then isotropically scaled to fit within a unit sphere centered at the origin. This step normalizes differences in absolute tooth size while preserving relative proportions.

Orientation preservation No rotational normalization is applied. The orientation of each tooth is preserved exactly as inherited from the original intra-oral scan. This design choice ensures that directional information relevant for bracket placement, such as bucco-lingual and mesio-distal directions, is retained and must be inferred by the model. Figure 4.4 provides a visual overview of various tooth instances extracted for the Melted dataset.

Metadata transformation The rigid transformation resulting from the centering and scaling operations is also applied to all associated orthodontic annotations. In particular, the landmark points of interest and the bracket base plane parameters are transformed into the same normalized reference frame as the tooth mesh. For each tooth instance, a dedicated JSON file is therefore generated, containing all annotation data expressed consistently with the normalized geometry.

The resulting Melted dataset consists of a collection of cleaned, normalized, and orientation-aware single-tooth meshes, each paired with an annotation file in a common reference frame. This representation enables efficient and precise training of the bracket placement model without the overhead of full-arch processing during learning, while preserving the information required for accurate and anatomically consistent landmark prediction.



(a) Incisor (b) Canine (c) Premolar (d) Molar

Figure 4.4: Actual mesh samples extracted from the Melted dataset.

5. Models

5.1 Introduction

This chapter describes the learning-based models developed for automatic dental bracket placement. The proposed pipeline is composed of two main components: (i) a semantic segmentation model for intra-oral scans, responsible for isolating individual teeth from raw point clouds, and (ii) a bracket positioning model, which predicts the optimal placement parameters for orthodontic brackets on each tooth surface.

5.2 Segmentation Model

5.2.1 Model Structure

The first component of the proposed pipeline is a semantic segmentation model designed to parse intra-oral scans of dental arches. Given an input point cloud representing either an upper or a lower arch, the model predicts a per-point semantic label corresponding to individual teeth and background regions. This step is fundamental, as accurate tooth isolation is a prerequisite for the subsequent bracket positioning model.

The segmentation network is based on the *PointTransformerV3* architecture, adopting an encoder–decoder design tailored for large-scale point cloud processing. The encoder progressively extracts hierarchical geometric features through self-attention mechanisms operating directly on point sets, while the decoder reconstructs point-wise predictions at the original resolution. This architecture allows the model to effectively capture both local tooth morphology and global dental arch structure.

The model is trained on the segmentation dataset described in Chapter 4.3. Instead of relying on manually annotated labels, the ground-truth segmentation maps are generated using the predictions of the ToothInstanceNet model. This strategy can be interpreted as a form of knowledge distillation, where a complex but accurate teacher model transfers its knowledge to a lighter and more efficient

student network. The motivation behind this choice lies in the computational characteristics of ToothInstanceNet: although highly accurate, it relies on a multi-stage processing pipeline that makes it unsuitable for real-time or large-scale deployment.

By contrast, our PointTransformerV3-based implementation, developed using the *Pointcept* framework [8], provides a substantial speedup while benefiting from a standardized and modular codebase. This makes the proposed segmentation model more suitable for integration in a production environment, without significantly compromising accuracy.

Qualitative evaluations show that the resulting segmentation quality is sufficient for the downstream task of bracket positioning. Consequently, this model is deployed in production to automatically isolate individual tooth meshes from intra-oral scans, serving as the input to the bracket placement prediction model described in the next section.

5.2.2 Implementation Details

Network architecture. The semantic segmentation model is based on a PointTransformerV3 encoder–decoder architecture. The network operates directly on 3D point clouds represented by their Cartesian coordinates. The encoder comprises five hierarchical stages with depths $(2, 2, 2, 6, 2)$ and channel dimensions $(32, 64, 128, 256, 512)$, while the decoder restores spatial resolution through four stages with channels $(64, 64, 128, 256)$. Multi-head self-attention is employed throughout the network to capture both local and global geometric context.

Data augmentation. Input scans are voxelized using a grid size of 0.01 to enforce a uniform point density. This step is required by the PointTransformer architecture, which relies on voxelization to compute positional embeddings through sparse convolution operations. The model predicts per-point semantic labels over 17 classes, corresponding to 16 teeth following the FDI notation and a gingiva class. During training, data augmentation is applied online and consists of small random rotations around the three Cartesian axes (± 0.1 radians around the x, y, and z axes), random isotropic scaling in the range $[0.9, 1.1]$, and random translations independently sampled along each axis within ± 0.05. No test-time augmentation is applied during inference.

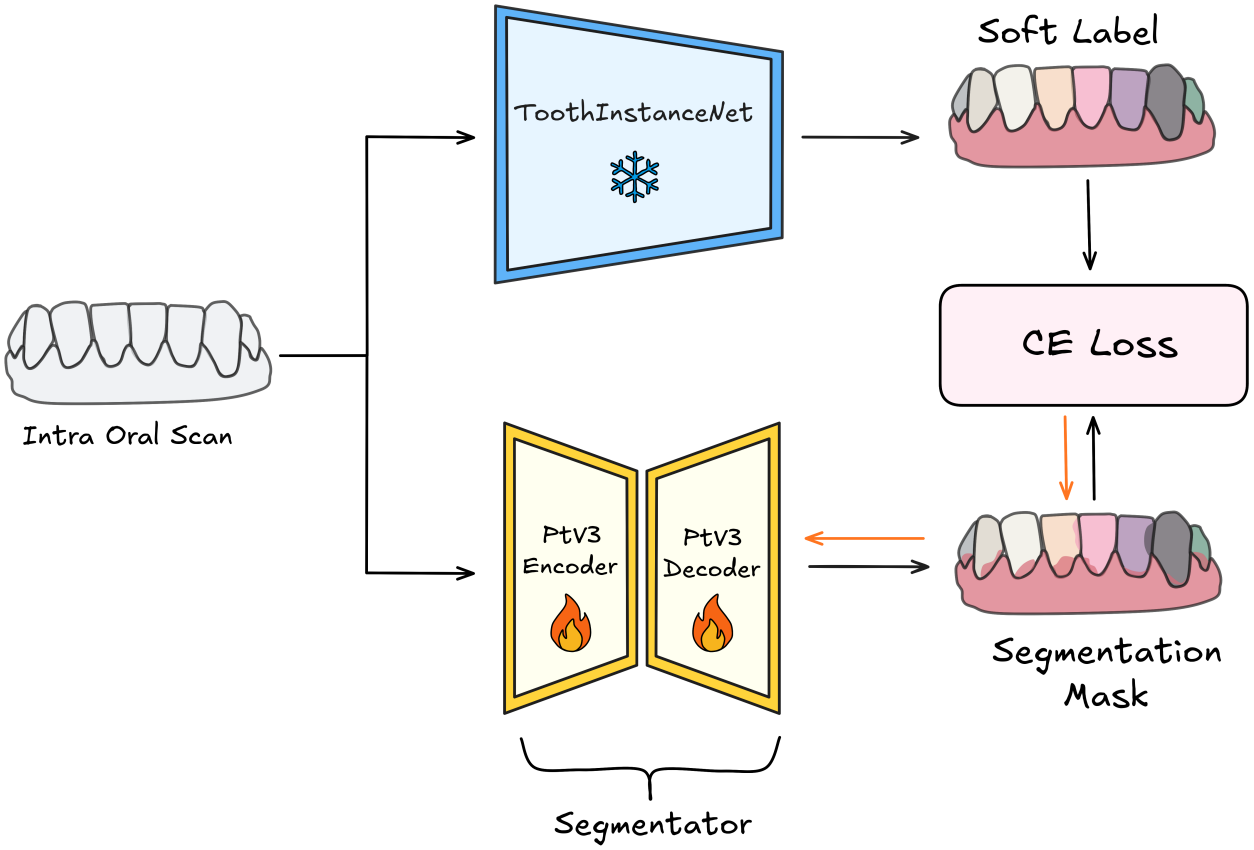


Figure 5.1: Training procedure of the proposed intra-oral scan segmentation model. ToothInstanceNet is used to generate soft labels for the training samples, with its weights kept frozen throughout the process. The segmentation network is trained to reproduce these predictions, yielding a lighter and more efficient model suitable for production deployment.

Training protocol. Training is conducted for 60 epochs with a batch size of 16, using the AdamW optimizer with a learning rate of 5×10^{-4} and a weight decay of 0.05. A One-Cycle learning rate policy with cosine annealing is applied, and gradient norms are clipped to 1.0. Mixed-precision training is enabled to improve computational efficiency. A standard Cross Entropy loss is used to guide the model during training. Figure 5.1 showcases the training pipeline of the segmentation model. The full training procedure took $\sim$ 12 hours on a single-GPU setup.

5.3 Bracket Position Prediction Model

The second component of the proposed pipeline is a model for predicting orthodontic bracket placement on individual tooth meshes. Given a segmented tooth represented as a point cloud, the model predicts three anatomically meaningful landmarks: the bracket installation point, the incisal reference point, and the gingival reference point. After landmark prediction, the local plane on which the bracket is placed is recovered through a simple geometric construction. The primary axis of the plane is defined by the direction connecting the gingival and incisal reference points, while the surface normal is estimated from the mesh at the predicted bracket location. The third orthogonal axis is then obtained as the cross product between these two vectors, yielding a local reference frame for bracket placement.

To improve robustness, the surface normal is computed as a weighted average of the normal vectors in a local neighborhood around the predicted bracket point, rather than from a single mesh face. This reduces sensitivity to local surface irregularities and prevents unstable plane estimation when predictions lie on regions of high curvature. The underlying deep learning architecture is again based on a PointTransformerV3 encoder and is instantiated in two different variants, which are described in the following sections.

5.3.1 Direct Regression Model

In the first variant, landmark localization is formulated as a direct regression problem. The PointTransformerV3 encoder extracts point-wise features, which are pooled and processed by three independent regression heads, one for each landmark. Each head predicts a 3D point $\hat{\mathbf{p}}_k \in \mathbb{R}^3$, with $k \in \{\text{bracket}, \text{incisal}, \text{gingival}\}$.

The model is trained using a standard ℓ_2 regression loss:

$$\mathcal{L}_{\text{reg}} = \sum_{k=1}^{3} \|\hat{\mathbf{p}}_k - \mathbf{p}_k\|_2^2 , \tag{5.1}$$

where $\mathbf{p}_k$ denotes the ground-truth position of landmark k. While this formulation yields reasonable localization accuracy, it is sensitive to local geometric ambiguities and provides limited interpretability, motivating the alternative heatmap-based formulation described below.

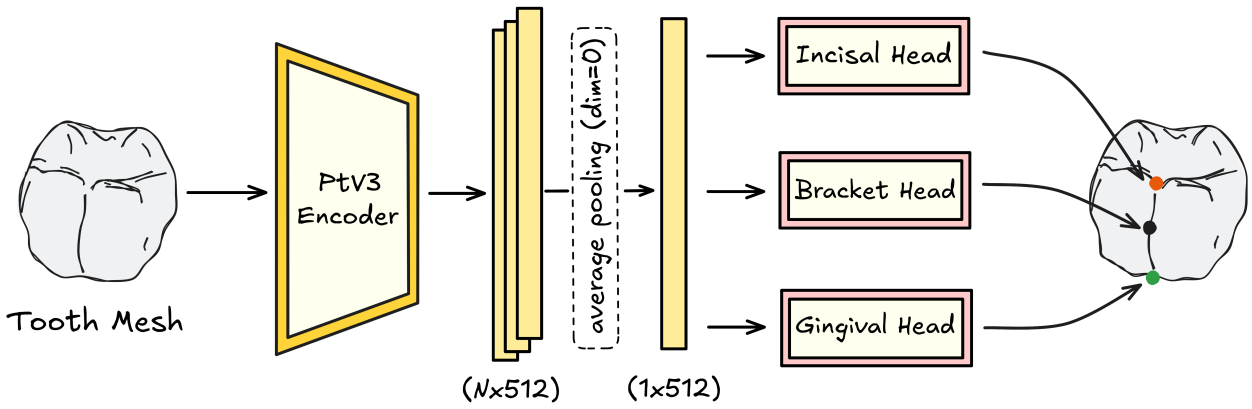


Figure 5.2: Bracket prediction model, simple regression variant.

5.3.2 Heatmap-Based Model

The second variant reformulates landmark localization as a dense per-vertex regression problem on the tooth surface. A PointTransformerV3 encoder–decoder architecture is employed to produce point-wise features at the original mesh resolution. A linear regression head maps these features to a three-channel output $\hat{\mathbf{H}} \in \mathbb{R}^{N\times 3}$, where N is the number of vertices in the tooth mesh and each channel corresponds to one landmark of interest.

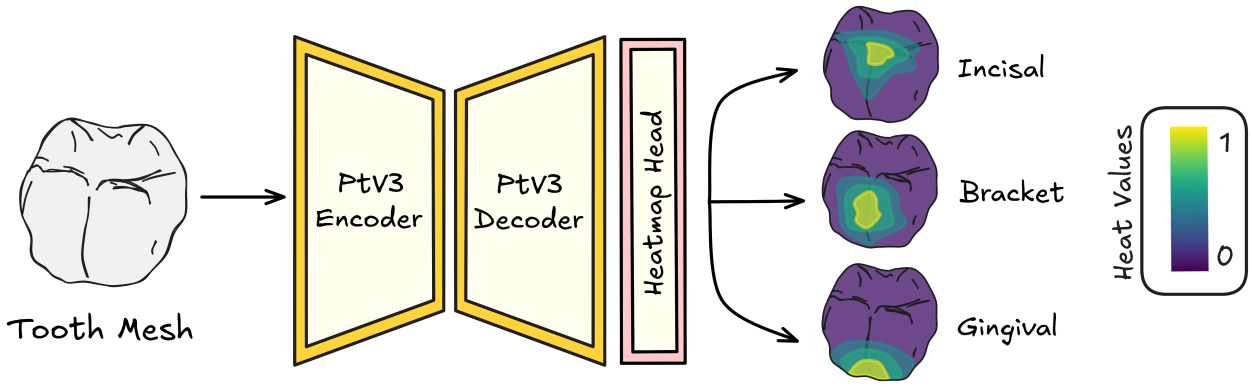


Figure 5.3: Bracket prediction model, heatmap variant.

Ground-truth heatmaps are constructed from the annotated dataset described in Section 4.2. For each landmark k and each vertex v_i, the target heatmap value is computed by applying a Gaussian kernel to the geodesic distance between the vertex and the annotated landmark position:

$$H_k(v_i) = \exp\left(-\frac{d_g(v_i, \mathbf{p}_k)^2}{2\sigma^2}\right), \tag{5.2}$$

where $d_g(\cdot,\cdot)$ denotes the geodesic distance on the tooth surface and σ controls the spatial spread of the heatmap. In all experiments, $\sigma = 0.15$ was found to provide a suitable trade-off between localization sharpness and spatial smoothness. This construction yields continuous, anatomically consistent supervision signals defined directly on the tooth surface. The use of geodesic distance is particularly well suited in this context, as it intrinsically respects the surface topology of the tooth. In contrast to the Euclidean distance in $\mathbb{R}^3$, which measures straight-line proximity through the volume, the geodesic distance evaluates the shortest path constrained to the tooth surface. This distinction is critical for thin dental structures, such as incisors, where vertices located on opposite sides of the tooth may exhibit a small Euclidean distance despite being anatomically unrelated. In such cases, a Euclidean-based Gaussian would incorrectly assign high heatmap values to vertices on the lingual surface when the landmark lies on the labial side. By relying on geodesic distance, vertices separated by the tooth surface receive large distance values, effectively yielding near-zero heatmap responses on the opposite side of the tooth. As a result, the generated heatmaps remain anatomically consistent and better aligned with the true spatial distribution of clinically relevant landmarks.

Overall, this heatmap-based formulation yields improved robustness to geometric ambiguities, higher localization accuracy, and enhanced interpretability compared to direct coordinate regression, and is therefore adopted as the final bracket prediction model in the proposed system.

Training is performed by minimizing a per-vertex regression loss between the predicted and ground-truth heatmaps. Since each landmark channel is treated independently and no normalization across vertices or channels is enforced, the loss is defined as the mean squared absolute error (ℓ_2 loss),

$$\mathcal{L}_{\text{hm}} = \frac{1}{3N} \sum_{k=1}^{3} \sum_{i=1}^{N} \left\| \hat{H}_k(v_i) - H_k(v_i) \right\|_2^2 \tag{5.3}$$

where $\hat{H}(v_i)$ denotes the output of the model.

At inference time, landmark coordinates are extracted from the predicted heatmaps by selecting the vertices whose activation values lie above the 95th percentile for each landmark channel. The final landmark position is then computed as a probability-weighted average of the selected vertices:

$$\hat{\mathbf{p}}_k = \frac{\sum_{i \in \mathcal{S}_k} \hat{H}_k(v_i)\, \mathbf{x}_i}{\sum_{i \in \mathcal{S}_k} \hat{H}_k(v_i)}, \tag{5.4}$$

where $\mathbf{x}_i$ denotes the 3D coordinates of vertex v_i and $\mathcal{S}_k$ is the set of selected vertices for landmark k.

Figure 5.3 illustrates the overall workflow of the heatmap-based variant, while Figure 5.4 showcases the 3-channel heatmap prediction for an actual sample case extracted from the Melted dataset.

5.3.3 Implementation Details

The bracket position prediction model is implemented in PyTorch [28] using the Pointcept [8] framework, following a configuration and training protocol closely aligned with that of the segmentation model to ensure architectural consistency across the pipeline.

Network architecture. The model is built upon a PointTransformerV3 backbone (PT-v3m1). The network operates directly on voxelized point clouds representing individual teeth and uses only the Cartesian coordinates (x, y, z) as input features. The encoder follows a hierarchical design with five stages of depths $(2, 2, 2, 6, 2)$ and corresponding channel dimensions $(32, 64, 128, 256, 512)$, while the decoder, used in the heatmap-based variant, restores point-wise resolution through four stages with channels $(64, 64, 128, 256)$. Multi-head self-attention is employed throughout the network, with the number of attention heads increasing with feature dimensionality, enabling the model to capture both fine-grained surface details and global tooth morphology.

Data augmentation. During training, online data augmentation is applied to improve robustness to small variations in tooth pose and orientation. Augmentations include random rotations around the x, y and z-axes within ± 0.1 radians, random isotropic scaling in the range $[0.9, 1.1]$ and random translations independently sampled along each axis within ± 0.05. At test time, data augmentation is also employed to improve prediction consistency. For each sample, three forward passes are performed, each corresponding to a different randomly sampled rotation around the three axes, applied independently with a probability of 50% per axis and with rotation angles uniformly sampled in the range $[-0.1, 0.1]$ radians. The final prediction is obtained by averaging the outputs of the three augmented inference runs.

Training protocol. The model is trained for 80 epochs with a batch size of 16. Optimization is performed using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 5×10^{-3}. A One-Cycle learning rate schedule with cosine annealing is adopted, with a warm-up

phase covering the first 10% of training iterations. Gradient norms are clipped to 1.0 to stabilize training, and automatic mixed-precision (AMP) is enabled to reduce memory consumption and improve computational efficiency. Model training takes $\sim$ 2 hours on a single-GPU setup.

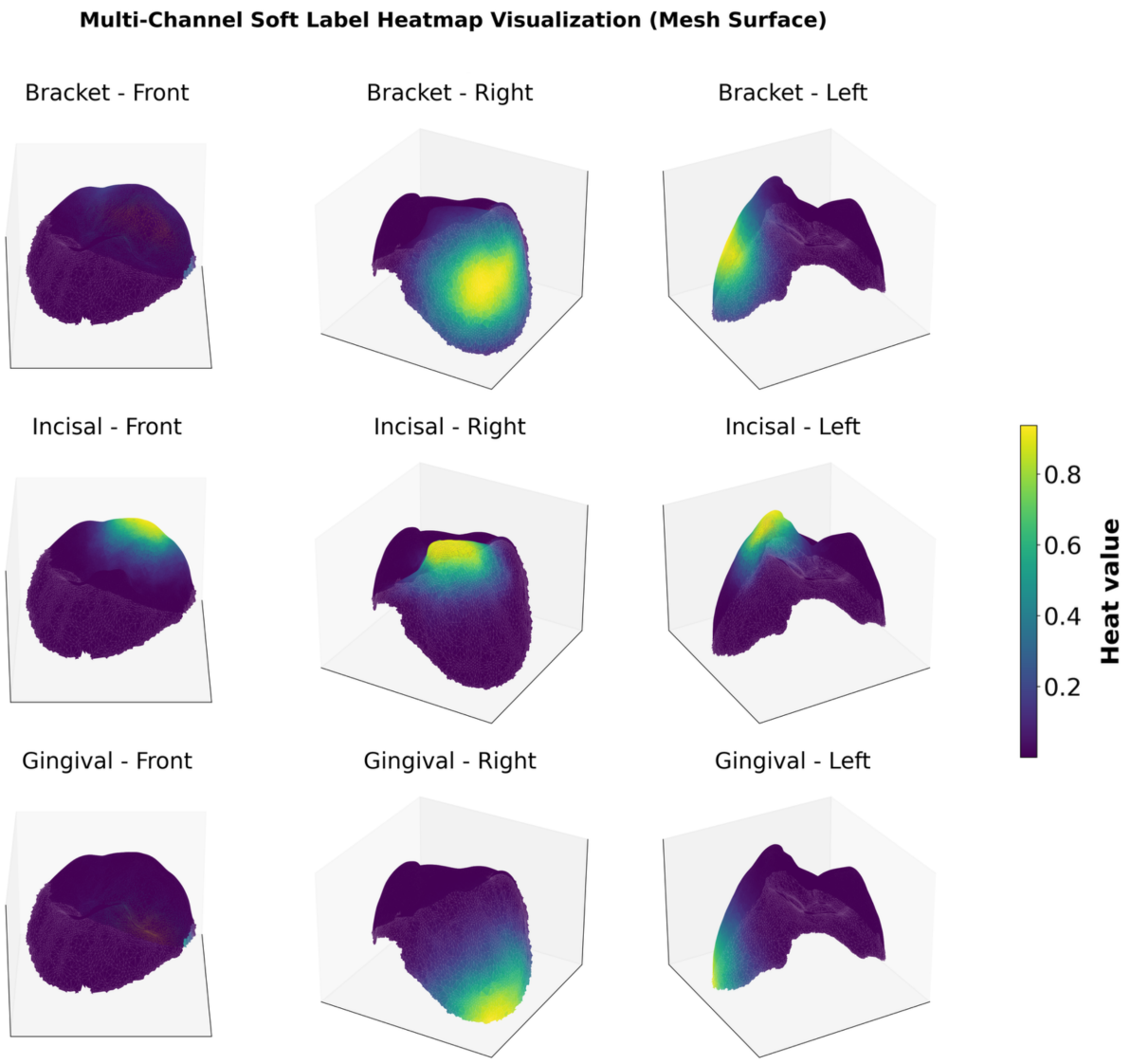


Figure 5.4: Multi-channel heatmap prediction on a sample premolar. For each channel, three complementary viewpoints are provided to enable a more accurate inspection of the tooth surface and its spatial structure. The heat value associated with each mesh face is computed, for visualization purposes, as the average of the soft-label values defined at its three vertices, resulting in a smooth and interpretable surface representation.

6. Experiments

6.1 Introduction

This chapter presents the experimental evaluation of the proposed framework for automatic orthodontic bracket placement on individual tooth meshes. The experiments aim to assess both the accuracy of the proposed learning-based models and the impact of key design choices, including output formulation, backbone architecture, and input feature representation.

A clinically motivated geometric baseline is first introduced to provide a deterministic reference for comparison. A per-tooth evaluation is then conducted to analyze performance across different tooth types and dental arches. Finally, a cross-validation study systematically compares heatmap-based and regression-based formulations, as well as different backbone architectures and input feature configurations.

All evaluations are performed using patient-level data splits to prevent information leakage. Performance is measured as the Euclidean distance in millimeters between predicted and clinically annotated bracket installation points.

6.1.1 Geometrical Baseline

To provide a clinically grounded reference for the evaluation of the proposed learning-based models, we introduce a geometrical baseline that closely reproduces the manual procedure adopted by orthodontic experts for bracket placement. The method is entirely deterministic after the extraction of anatomical landmarks with ToothInstanceNet, since it uses surface-aware geometric constructions on the tooth mesh.

Since ToothInstanceNet outputs multiple landmark proposals per tooth, each associated with a confidence score, the final landmark locations are obtained by computing a weighted average of the top 95^{th} percentile of proposals, using the predicted scores as weights. This aggregation strategy improves robustness to outliers and landmark detection noise.

For each tooth, the bracket placement procedure is performed independently and consists of the

following steps:

- **Mesio–distal reference construction.** The mesial and distal landmarks are connected by a geodesic curve constrained to the tooth surface. This curve defines the mesio–distal reference direction commonly used in clinical practice.
- **Initial incisal reference estimation.** The midpoint of the mesio–distal geodesic is computed and used as an initial approximation of the incisal reference point. This assumption is generally valid but may be affected by irregularities in the incisal ridge geometry.
- **Curvature-based refinement.** To better match the location from which orthodontists perform measurements, a local curvature estimation is applied to shift the incisal reference apically from the ridge crest toward the enamel edge. This refinement step was empirically observed to slightly reduce the overall bracket localization error.
- **Incisal–gingival geodesic construction.** A second geodesic curve is computed between the refined incisal reference point and the gingival landmark extracted by ToothInstanceNet. This curve represents the incisal–gingival measurement axis along the tooth surface.
- **Bracket position selection.** The geodesic distance between the incisal reference and the gingival landmark is measured in millimeters. Based on the tooth-specific clinical guidelines shown in Figure 2.5, the point along the incisal–gingival geodesic corresponding to the prescribed distance is selected as the final bracket installation point.

By relying exclusively on geodesic distances, this baseline ensures that all measurements are consistent with the tooth surface topology, avoiding artifacts that would arise from Euclidean distances in $\mathbb{R}^3$. While this approach faithfully reproduces expert-defined geometric reasoning, it remains sensitive to landmark inaccuracies and anatomical variability, as it lacks the adaptive capacity of learning-based models. For this reason, it is used exclusively as a reference baseline in the experimental evaluation.

6.1.2 Per-Tooth Comparison of Bracket Placement Accuracy Against Baseline Methods

This experiment evaluates the accuracy of bracket installation point prediction on the Melted dataset, described in section 4.4. The goal is twofold: (i) to assess the consistency of the proposed heatmap-based PointTransformerV3 model across different tooth types and dental arches, and (ii) to verify whether facial landmarks commonly available in public datasets are sufficiently aligned with clinically annotated bracket placement points.

Three different methods are evaluated and compared:

- **Heatmap-based PointTransformerV3 (proposed):** our learning-based model trained to directly regress a per-tooth heatmap encoding the bracket installation point, further explained in section 5.2
- **Geometrical baseline**, explained in the previous section 6.1.1.
- **Facial landmark baseline:** a baseline that directly uses the facial landmark predicted by ToothInstanceNet (see section 3.6) as the bracket installation point, without additional refinement.

Experimental Setup. The evaluation was conducted using a single-fold, patient-level data split, where 80% of the patients were used for the training of our model and the remaining 20% for testing. The split was strictly enforced at the patient level to prevent data leakage, ensuring that all teeth extracted from a given patient were assigned exclusively either to the training set or to the test set.

The heatmap-based PointTransformerV3 model was trained exclusively on the training set and evaluated on the held-out test set. ToothInstanceNet was not trained as part of this experiment and was employed only at inference time on the test set, both to extract reference points for the geometrical baseline and to obtain the facial landmark used as another baseline.

Quantitative results. A per-tooth breakdown of the prediction error is reported in Figure 6.1, where errors are grouped by FDI index. In addition, arch-level aggregated statistics are summarized in Table 6.1. The results indicate that the proposed heatmap-based PointTransformerV3 model consistently outperforms both baselines across all evaluation settings.

Table 6.1: Bracket installation point prediction error for the three evaluated methods, reported separately for the maxillary (upper) arch, mandibular (lower) arch, and overall test set. Values are given as mean ± standard deviation in mm.

Method	**Error (mm)**		
	Maxillary	**Mandibular**	**Overall**
ToothInstanceNet [26]	0.58 ± 0.33	0.79 ± 0.58	0.69 ± 0.49
Geometric baseline	0.66 ± 0.51	0.95 ± 0.63	0.81 ± 0.59
Heatmap PointTransformerV3 (ours)	**0.53 ± 0.55**	**0.53 ± 0.52**	**0.53 ± 0.53**

For the upper arch, the proposed model achieves a mean error of 0.53±0.55 mm, improving upon both the landmark-based method (0.58 ± 0.33 mm) and the geometric baseline (0.66 ± 0.51 mm). The improvement is more pronounced in the lower arch, where the proposed approach reduces the mean error to 0.53 ± 0.52 mm, compared to 0.78 ± 0.57 mm for the ToothInstanceNet landmark and 0.95 ± 0.63 mm for the geometric baseline. Overall, across both arches, the proposed model achieves an average error of 0.53 (roughly half millimeter) over 741 teeth.

Discussion. The results demonstrate that the proposed model not only improves average prediction accuracy but also maintains stable performance across different tooth types and anatomical regions. In contrast, the facial landmark extracted from ToothInstanceNet exhibits a notable degradation in the mandibular arch, suggesting a mismatch between the generic facial landmark definition commonly found in public datasets and the clinically relevant bracket placement point considered in this work. By directly learning a task-specific heatmap representation, the proposed PointTransformerV3 model better captures the orthodontic annotation protocol, resulting in consistent improvements across the entire dental arch.

6.1.3 Cross-Validation Study of Model Variants

This experiment is designed to isolate and quantify the impact of architectural choices and input feature representations on bracket installation point prediction accuracy. In particular, we compare heatmap-based and regression-based formulations, as well as the effect of incorporating surface

normal information in addition to 3D point coordinates.

Cross-validation protocol. All models are evaluated using a 5-fold cross-validation scheme defined at the patient level. For each fold, patients are randomly partitioned into three disjoint subsets: (i) approximately 20% of the patients are held out as a test set, (ii) approximately 72% are used for training, and (iii) the remaining 8% are used for validation. Minor variations in these percentages arise due to the indivisibility of patient samples across folds.

The patient-level split is strictly enforced to prevent information leakage, ensuring that no tooth from a given patient appears in more than one subset. For each fold, models are trained from scratch on the corresponding training set, early stopping is performed according to the performance on the validation set, and final performance is reported on the held-out test set. The results presented below are obtained by aggregating the test-set metrics across the five folds.

Compared models. We evaluate a total of eight model variants by implementing four distinct task configurations across two different architectural backbones: PointTransformerV3 and a sparse convolution-based implementation of the Unet, that we call Sp-Unet. Each configuration is tested with both backbones to assess the impact of architecture on performance. The configurations are defined as follows:

- **Heatmap (coordinates only).** The heatmap-based model described in section 5.3.2 taking as input only the 3D coordinates of the tooth point cloud.
- **Heatmap (coordinates + normals).** The same heatmap-based architecture augmented with per-point surface normal vectors as additional input features.
- **Regression (coordinates only).** The regression-based variant explained in section 5.3.1 which leverages point cloud coordinates alone.
- **Regression (coordinates + normals).** The regression-based with normal vectors as additional input features.

All models share the same backbone architecture and training protocol, differing only in their output formulation (heatmap vs. regression) and input feature dimensionality. This controlled setup allows for a fair comparison of the respective design choices.

Backbone	Model	Error (mm)		
		Maxillary	**Mandibular**	**Overall**
PTV3 [39]	Heatmap (coordinates)	**0.48 ± 0.34**	**0.55 ± 0.54**	**0.51 ± 0.46**
	Heatmap (coordinates + normals)	0.48 ±0.36	0.54 ±0.58	0.51 ±0.50
	Regression (coordinates)	0.56 ±0.41	0.57 ±0.58	0.56 ±0.51
	Regression (coordinates + normals)	0.55 ±0.41	0.58 ±0.60	0.56 ±0.52
Sp-unet	Heatmap (coordinates)	0.53 ±0.46	0.57 ±0.61	0.55 ±0.54
	Heatmap (coordinates + normals)	0.53 ±0.44	0.56 ±0.62	0.55 ±0.54
	Regression (coordinates)	0.55 ±0.41	0.58 ±0.56	0.56 ±0.50
	Regression (coordinates + normals)	0.55 ±0.40	0.6 ±0.58	0.57 ±0.51

Table 6.2: Comparison of heatmap-based and regression-based models using different backbone architectures. Results are reported as mean Euclidean error (mm) ± standard deviation over a 5-fold cross-validation.

Evaluation metric. Performance is measured as the Euclidean distance, in millimeters, between the predicted bracket installation point and the clinically annotated ground-truth location. For regression-based models, which directly predict a point in $\mathbb{R}^3$, the final bracket position is obtained by projecting the predicted point onto the tooth surface, using the closest face on the mesh. This projection ensures that all predictions lie on the anatomical surface and are therefore comparable to the ground-truth annotations. Results are reported as mean error ± standard deviation across all teeth in the test sets of the five folds.

Quantitative results. Table 6.2 summarizes the cross-validation results for the four evaluated models. The table reports the average performance across folds and can be directly compared to the single-fold results presented in the previous section.

Discussion. The cross-validation results reported in Table 6.2 highlight clear trends regarding the output formulation, backbone architecture, and input feature representation. Overall, the heatmap-based variant using only 3D point coordinates and the PointTransformerV3 backbone achieves the

best performance across all evaluation settings, obtaining the lowest mean error for both maxillary and mandibular arches. This result confirms the effectiveness of dense per-vertex likelihood modeling for orthodontic bracket localization and demonstrates the strong representational capacity of PointTransformerV3 in this formulation.

In contrast, the regression-based variants do not exhibit a comparable performance gain from the use of the PointTransformerV3 backbone when compared to the Sp-UNet architecture. For direct coordinate regression, both backbones yield similar error distributions, suggesting that the increased modeling capacity of PointTransformerV3 does not translate into improved accuracy for this output formulation. This indicates that the benefits of more expressive point-based attention mechanisms are best exploited when coupled with a dense prediction objective rather than direct point regression.

Across both backbones and output formulations, the inclusion of surface normal vectors does not lead to consistent performance improvements. In several cases, coordinate-only models match or slightly outperform their coordinate-plus-normal counterparts. This suggests that explicit normal information introduces additional input complexity without providing meaningful gains, as the models are able to infer the relevant geometric structure of the tooth surface directly from the spatial arrangement of point coordinates.

Taken together, these results indicate that accurate bracket placement prediction can be achieved using minimal geometric input features, provided that an appropriate dense prediction formulation and backbone architecture are employed. The heatmap-based model with PointTransformerV3 and coordinate-only input therefore represents the most effective and robust design choice among the evaluated configurations.

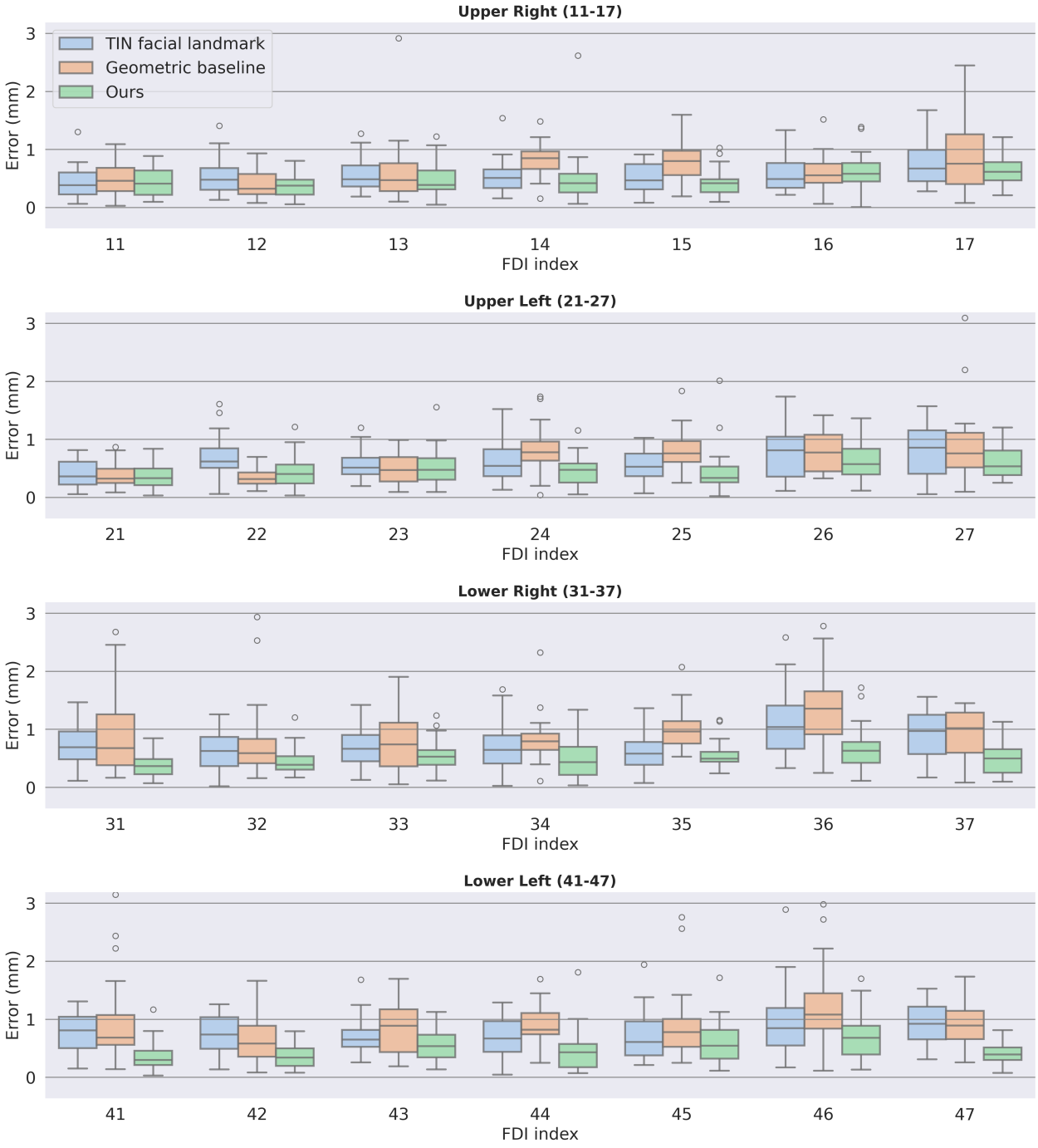


Figure 6.1: Per-tooth bracket positioning error (mm) indexed by FDI notation for three different bracket localization methods: (1) the geometric baseline, (2) the facial landmark extracted from ToothInstanceNet predictions, and (3) the proposed heatmap-based PointTransformerV3 model. Results are shown per tooth type, demonstrating consistent performance across dental anatomies.

7. Autobonding

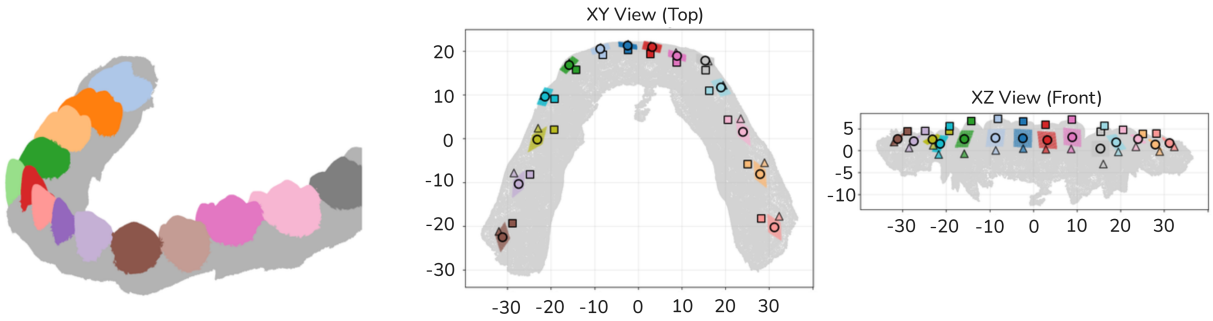
The Autobonding system represents the web-accessible interface developed to deploy the orthodontic bracket placement model in a manner that is directly integrable with existing clinical software. By exposing the segmentation and prediction pipeline through a RESTful API, Autobonding facilitates seamless access for orthodontic specialists, enabling them to utilize the predictive capabilities of the model without requiring direct interaction with the underlying machine learning framework or computational infrastructure.

Autobonding is deployed within a containerized environment using Docker [22]. This design choice ensures that the entire pipeline, including all dependencies, libraries, and model weights, can be executed reliably across different computing environments. Dockerization simplifies maintenance, deployment, and potential local execution, providing a standardized and reproducible setup.

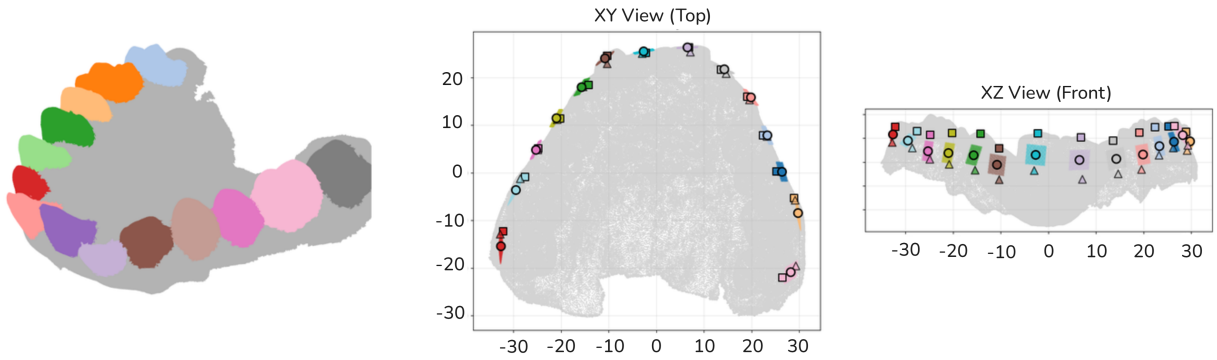
The service is hosted on dedicated servers equipped with an NVIDIA 2080 Ti graphics card. The model is persistently loaded on the GPU, which allows for low-latency inference. From the moment a prediction request is received to the point at which the results are returned, the system exhibits response times ranging between 10 and 15 seconds, depending on the size of the intra-oral scan. This performance ensures that predictions can be integrated effectively into clinical workflows without introducing significant delays.

To ensure high predictive reliability, Autobonding generates and monitors a comprehensive suite of evaluation plots for every prediction, shown in Figure 7.1. These visualizations offer critical insights into segmentation quality and bracket placement accuracy, enabling both developers and clinical collaborators to evaluate model performance efficiently. Furthermore, an administrative dashboard (see Figure 7.2) facilitates the management of bonding requests and the inspection of generated analytical plots. Final outputs are delivered to the user via a JSON file containing the specific bracket coordinates for the uploaded intra-oral scan.

Autobonding also incorporates a feedback loop that allows orthodontic specialists to submit corrections and additional annotations directly through the platform. This capability enables continuous refinement of the model over time, integrating expert knowledge into future iterations. By



(a) Mandibular (lower)



(b) Maxillary (upper)

Figure 7.1: Dashboard visualization of model outputs for a complete clinical case, showing segmented dental structures and predicted bonding sites for the mandibular (a) and maxillary (b) dental arches.

systematically collecting and incorporating these corrections, the system can improve prediction accuracy and robustness, gradually adapting to a wider variety of patient anatomies and clinical scenarios.

The Autobonding API represents a practical solution for integrating advanced machine learning models into real-world orthodontic workflows. By combining containerized deployment, GPU-accelerated inference, quality monitoring, and expert-driven continuous refinement, the system provides a reliable, maintainable, and clinically useful tool for automated bracket placement.

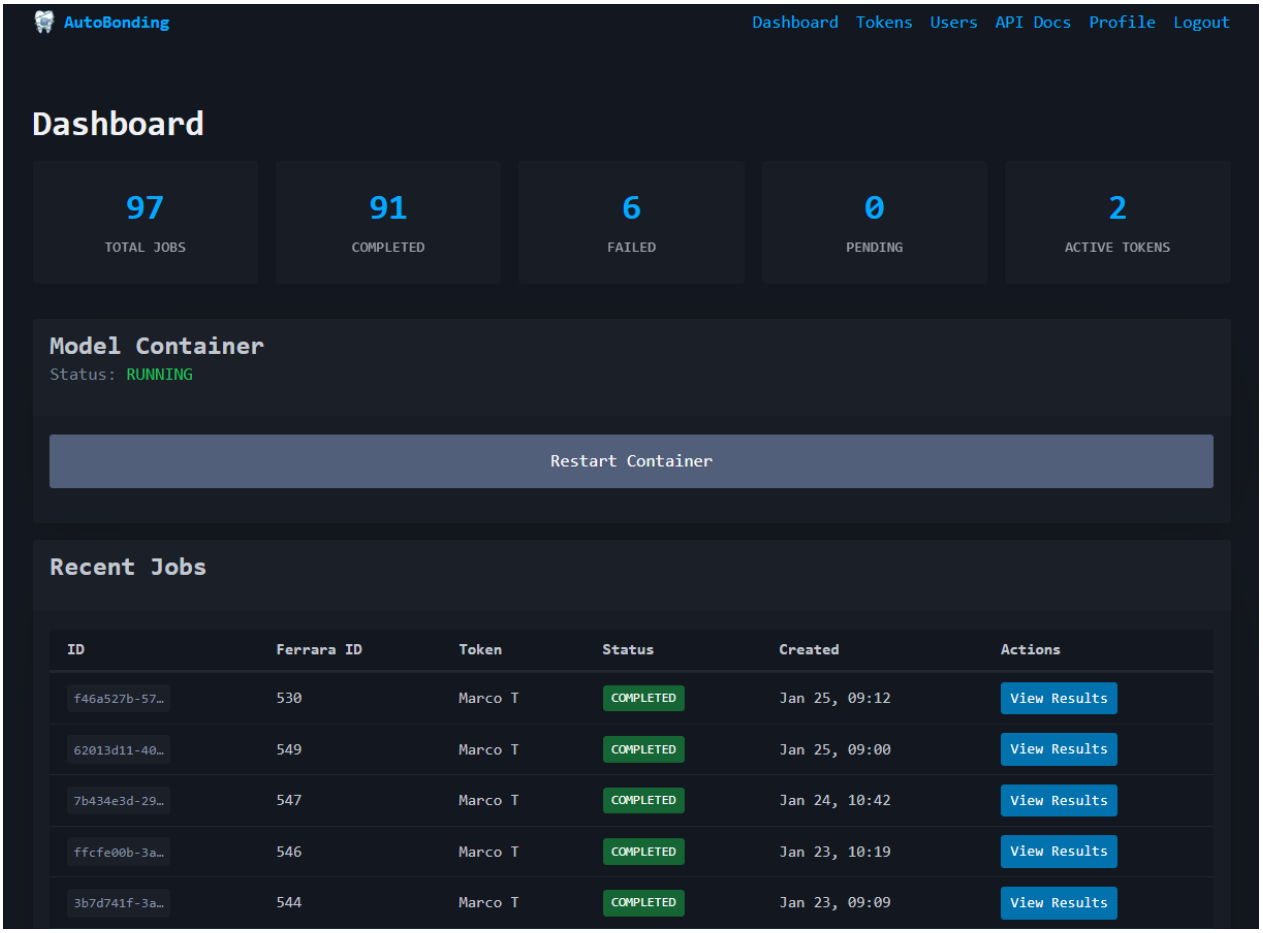


Figure 7.2: The Autobonding dashboard.

8. Conclusions and Future Work

This thesis presented a learning-based framework for the automatic prediction of orthodontic bracket installation points starting from intra-oral scans. The proposed approach addresses a clinically relevant and time-consuming task by leveraging 3D deep learning models operating directly on dental geometries, with the objective of improving accuracy, reproducibility, and integration within digital orthodontic workflows.

The work covered the full pipeline required for practical deployment, including data preparation, tooth-level segmentation, geometric normalization, prediction of bracket placement points, and exposure of the system through a dedicated web API. A key contribution of this thesis is the design and evaluation of multiple prediction strategies, ranging from geometrical baselines inspired by clinical practice to learning-based models exploiting both local tooth geometry and global contextual information. Experimental results demonstrated that the proposed models achieve consistent and clinically meaningful predictions on the available dataset.

Another important contribution is the introduction of the Melted dataset, a representation that enables aggregation and analysis of tooth shapes across patients and classes. This formulation proved effective for training robust models while reducing variability introduced by individual anatomical differences. In addition, the development of a custom segmentation model and its integration into a containerized deployment pipeline ensured flexibility, maintainability, and real-world applicability of the system.

Overall, this thesis shows that automatic bracket placement prediction from intra-oral scans is a feasible and promising direction, and that learning-based approaches can effectively model the complex geometric relationships underlying orthodontic expertise.

While the proposed system demonstrates encouraging results, several directions remain open for future research and development.

First, model accuracy can be further improved by incorporating additional corrective mechanisms and richer supervision. In particular, integrating systematic correction steps on the predicted bracket positions and expanding the training dataset with additional annotations provided by orthodontic specialists would allow the model to better capture clinical variability and expert prefer-

ences. Increased annotation diversity is expected to improve both robustness and generalization.

Second, although a custom tooth segmentation model was developed and deployed to meet performance and flexibility requirements, its effectiveness has not yet been quantitatively evaluated against state-of-the-art segmentation methods. A comparison with existing approaches on standard benchmarks would provide a clearer understanding of its strengths and limitations, and could guide further architectural or training improvements.

Third, the Melted dataset introduced in this work represents a valuable resource that has not yet been fully exploited. Beyond its use for bracket placement prediction, it could be leveraged to model canonical tooth shapes for different classes, enabling more compact shape representations and potentially improving both segmentation and prediction tasks. Exploring generative or statistical shape modeling techniques on this dataset is a promising direction.

Finally, future work will focus on further reducing user interaction by developing an additional module capable of automatically orienting incoming intra-oral scans. Such a component would remove the need for manual alignment by the orthodontic specialist, improving usability and enabling a fully automated end-to-end pipeline from raw scan acquisition to bracket placement prediction.

In conclusion, this thesis lays the foundation for an integrated, learning-based system for automatic orthodontic bracket placement. The proposed future extensions aim to improve accuracy, usability, and clinical relevance, bringing the system closer to real-world adoption in digital orthodontics.

Acknowledgments

Ringrazio innanzitutto Fede Bolelli e Luca, che mi hanno accompagnato in questo percorso e dai quali ho imparato molto. Mi sono divertito durante il tirocinio e sono conteto di aver trovato il lab. Zip!

Grazie a mamma e papà, che mi avete dato l'opportunità di studiare. Spero che, con il tempo, si comprenda quanto questo sia un privilegio tutt'altro che scontato.

Grazie, Marti. Concludiamo insieme un viaggio che, in fondo, abbiamo percorso fianco a fianco. Lungo il cammino sono successe tante cose, ma ci siamo sempre sostenuti e, alla fine, abbiamo raggiunto questo traguardo insieme.

Infine ringrazio il prof. Federico Lugli, Dotti, Enri, Cabro, Carlo, gli amici della parrocchia di Budrione, il Decumano Ovest e la Contea APS!

Bibliography

[1] Pokpong Amornvit, Dinesh Rokaya, and Sasiwimol Sanohkan. “Comparison of accuracy of current ten intraoral scanners”. In: *BioMed research international* 2021.1 (2021), p. 2673040.

[2] Nasib Balut et al. “Variations in bracket placement in the preadjusted orthodontic appliance”. In: *American journal of orthodontics and dentofacial orthopedics* 102.1 (1992), pp. 62–67.

[3] Achraf Ben-Hamadou et al. “Teeth3DS+: an extended benchmark for intraoral 3D scans analysis”. In: *arXiv preprint arXiv:2210.06094* (2022).

[4] Federico Bolelli et al. “Segmenting Maxillofacial Structures in CBCT Volumes”. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 5238–5248.

[5] Michael M Bronstein et al. “Geometric deep learning: going beyond euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42.

[6] Jieneng Chen et al. *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*. 2021. arXiv: 2102.04306 [cs.CV]. URL: https://arxiv.org/abs/2102.04306.

[7] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *CoRR* abs/1606.00915 (2016). arXiv: 1606.00915. URL: http://arxiv.org/abs/1606.00915.

[8] Pointcept Contributors. *Pointcept: A Codebase for Point Cloud Perception Research*. https://github.com/Pointcept/Pointcept. 2023.

[9] Alexey Dosovitskiy. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[10] Maxime Gillot et al. “Automatic landmark identification in cone-beam computed tomography”. In: *Orthodontics & craniofacial research* 26.4 (2023), pp. 560–567.

[11] Thorsten Grünheid, Shawn D. McCarthy, and Brent E. Larson. "Clinical use of a direct chairside oral scanner: An assessment of accuracy, time, and patient acceptance". In: *American Journal of Orthodontics and Dentofacial Orthopedics* 146.5 (2014), pp. 673–682. ISSN: 0889-5406. DOI: https://doi.org/10.1016/j.ajodo.2014.07.023. URL: https://www.sciencedirect.com/science/article/pii/S0889540614007276.

[12] Albert Gu and Tri Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. 2024. arXiv: 2312.00752 [cs.LG]. URL: https://arxiv.org/abs/2312.00752.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[14] Yue Lai et al. "Automated Dental Landmarks Recognition in Special Models via Neural Network". In: *IEEE Transactions on Computational Social Systems* (2025).

[15] Juncheng Li et al. "A fine-grained orthodontics segmentation model for 3D intraoral scan data". In: *Computers in biology and medicine* 168 (2024), p. 107821.

[16] Pengcheng Li et al. "THISNet: Tooth Instance Segmentation on 3D Dental Models via Highlighting Tooth Regions". In: *IEEE Transactions on Circuits and Systems for Video Technology* 34.7 (2023), pp. 5229–5241.

[17] Xiang Li. *3D multimodal dental dataset based on CBCT and oral scan.* Sept. 2024. DOI: 10.6084/m9.figshare.26965903.v3. URL: https://figshare.com/articles/dataset/_b_3D_multimodal_dental_dataset_based_on_CBCT_and_oral_scan_b_/26965903.

[18] Chunfeng Lian et al. "Deep Multi-Scale Mesh Feature Learning for Automated Labeling of Raw Dental Surfaces From 3D Intraoral Scanners". In: *IEEE Transactions on Medical Imaging* 39.7 (2020), pp. 2440–2450. DOI: 10.1109/TMI.2020.2971730.

[19] Roberta Lione et al. "Accuracy, time, and comfort of different intraoral scanners: An in vivo comparison study". In: *Applied Sciences* 14.17 (2024), p. 7731.

[20] Baoyuan Liu et al. "Sparse convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 806–814.

[21] Xinhai Liu et al. “Spu-net: Self-supervised point cloud upsampling by coarse-to-fine reconstruction with self-projection optimization”. In: *IEEE Transactions on Image Processing* 31 (2022), pp. 4213–4226.

[22] Dirk Merkel. “Docker: lightweight linux containers for consistent development and deployment”. In: *Linux journal* 2014.239 (2014), p. 2.

[23] MICCAI 2024. *3dTeethLand challenge*. Accessed 01-17-2026. 2026. URL: https://www.synapse.org/Synapse:syn57400900/wiki/627259.

[24] Bongki Moon et al. “Analysis of the clustering properties of the Hilbert space-filling curve”. In: *IEEE Transactions on knowledge and data engineering* 13.1 (2001), pp. 124–141.

[25] Tung Nguyen and Tate Jackson. “3D technologies for precision in orthodontics”. In: *Seminars in Orthodontics* 24.4 (2018). Digital Technologies In Orthodontics – An update, pp. 386–392. ISSN: 1073-8746. DOI: https://doi.org/10.1053/j.sodo.2018.10.003. URL: https://www.sciencedirect.com/science/article/pii/S1073874618300616.

[26] Niels van Nistelrooij and Shankeeth Vinayahalingam. “ToothInstanceNet: Comprehensive Information from Intra-oral Scans by Integration of Large-Context and High-Resolution Predictions”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 229–240.

[27] Nearchos C. Panayi et al. “Digital orthodontics: Present and future”. In: *AJO-DO Clinical Companion* 4.1 (2024), pp. 14–25. ISSN: 2666-4305. DOI: https://doi.org/10.1016/j.xaor.2023.12.001. URL: https://www.sciencedirect.com/science/article/pii/S2666430523001528.

[28] Adam Paszke et al. “Automatic differentiation in PyTorch”. In: (2017).

[29] Charles R Qi et al. “Pointnet: Deep learning on point sets for 3d classification and segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[30] Charles Ruizhongtai Qi et al. “Pointnet++: Deep hierarchical feature learning on point sets in a metric space”. In: *Advances in neural information processing systems* 30 (2017).

[31] Raphaël Richert et al. “Intraoral scanner technologies: a review to make a successful impression”. In: *Journal of healthcare engineering* 2017.1 (2017), p. 8427595.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[33] Abdelrahman Shaker et al. *UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation*. 2024. arXiv: 2212.04497 [cs.CV]. URL: https://arxiv.org/abs/2212.04497.

[34] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[35] Shankeeth Vinayahalingam et al. “Intra-oral scan segmentation using deep learning”. In: *BMC Oral Health* 23.1 (2023), p. 643.

[36] Xiaotong Wang et al. “Convolutional neural network for automated tooth segmentation on intraoral scans”. In: *BMC Oral Health* 24.1 (2024), p. 804.

[37] Brénainn Woodsend et al. “Automatic recognition of landmarks on digital dental models”. In: *Computers in Biology and Medicine* 137 (2021), p. 104819. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104819. URL: https://www.sciencedirect.com/science/article/pii/S0010482521006132.

[38] Xiaoyang Wu et al. “Point transformer v2: Grouped vector attention and partition-based pooling”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33330–33342.

[39] Xiaoyang Wu et al. “Point transformer v3: Simpler faster stronger”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 4840–4851.

[40] Zhirong Wu et al. “3d shapenets: A deep representation for volumetric shapes”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1912–1920.

[41] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090.

[42] Farhad Ghazvinian Zanjani et al. “Deep learning approach to semantic segmentation in 3D point cloud intra-oral scans of teeth”. In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2019, pp. 557–571.

[43] Hengshuang Zhao et al. “Point transformer”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 16259–16268.

[44] Yue Zhao et al. “Two-stream graph convolutional network for intra-oral scanner image segmentation”. In: *IEEE Transactions on Medical Imaging* 41.4 (2021), pp. 826–835.