A Comparative Study of 2D and 3D Deep Learning Approaches for Tooth Presence Classification from Intraoral Scans

Master Project

Lihui Zhou, Ruiqi Hong, Brighton Thomas

24-744-807, 23-756-786, 24-744-708

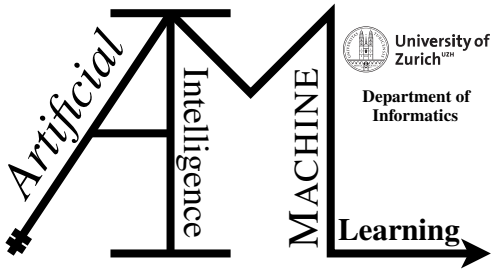
Submitted on

Janurary 29, 2026

Project Supervisor

Prof. Dr. Manuel Günther

Mr. Salman Mohammad



Declaration of Independence for Written Work

I hereby declare that I have **composed** this work independently and without the use of any aids other than those declared (including generative AI such as ChatGPT) – the use of generative AI to **improve** my composed work was permitted by the project supervisor. I am aware that I take full responsibility for the scientific character of the submitted text myself, even if AI aids were used. All passages taken verbatim or in sense from published or unpublished writings are identified as such. The work has not yet been submitted in the same or similar form or in excerpts as part of another examination.

Zürich, 29. 01. 2026
Place, Date

Lihui Zhou Ruiqi Hong Brighton Thomas
Lihui Zhou, Ruiqi Hong, Brighton Thomas

Master Project

Author: Lihui Zhou, Ruiqi Hong, Brighton Thomas,

Project period: January 29, 2025 - January 29, 2026

Artificial Intelligence and Machine Learning Group
Department of Informatics, University of Zurich

Acknowledgements

We are deeply grateful to have had the opportunity to work together as a team on this master project. Throughout this journey, we have overcome numerous challenges, and each team member has contributed to advancing this research. We would like to express our sincere gratitude to our supervisor, Prof. Dr. Manuel Günther, for his constructive feedback and for his excellent deep learning course, which introduced us to the field and laid the foundation for this project. We also extend special thanks to our mentor, Salman Mohammad, who guided us along the right path, provided invaluable insights, and was always available to answer our questions whenever we needed support. This experience has taught us that meaningful research emerges not just from individual effort but also from the collective dedication of those who believe in a shared vision.

Abstract

Intraoral scanning (IOS) technology provides precise 3D dental models. However, extracting clinical information, such as tooth presence for complete dental charting, still largely relies on manual interpretation.

This project investigates automated tooth presence classification under a realistic small-data setting with severe class imbalance, where missing teeth (especially in the anterior region) are rare. We develop and compare two deep learning pipelines: (i) a 2D pathway that renders 3D meshes into images and performs image-based multi-label classification, and (ii) a 3D pathway that directly processes point clouds derived from the meshes. To mitigate the long-tail label distribution, we introduce a hybrid, distribution-aware augmentation strategy that synthetically removes teeth from complete scans using tiered sampling probabilities and targeted pattern generation, and we further adopt the Dynamit loss to emphasize minority (missing-tooth) signals during training. Experiments on a render-matched held-out test set show a clear modality gap. While the 2D pathway can achieve strong performance on high-support posterior teeth, it is sensitive to rendering alignment and exhibits an anterior collapse under unconstrained scan orientations. In contrast, the 3D pathway is more robust to orientation variability and domain shifts, achieving better recall and balanced accuracy for rare missing-tooth detection. In general, our results support the processing of IOS data in its native 3D representation when reliable fine-grained dental charting is required under data scarcity.

Contents

Chapter 1

Introduction

1.1 Background and Significance

Dentistry is undergoing a rapid digital transformation. Traditional diagnosis methods now use data-driven technologies, especially 3D imaging. Intraoral scanning (IOS) is particularly important because it creates precise digital models of teeth without using radiation. More dental practices are adopting intraoral scanners due to improvements in accuracy, speed, and workflow Afrashthefar et al. (2022). These digital models now support virtual surgical simulations, orthodontic planning, and prosthetic predictions.

However, manual dental charting is still slow and prone to errors. Dentists must visually examine and document each tooth's condition and any restorations. This process varies between observers and lacks consistent notation standards. Deep learning has the potential to automate these complex tasks using geometric and textural information in IOS data Tan et al. (2025). Most existing research uses Cone-Beam Computed Tomography (CBCT) data, but IOS offers distinct advantages: it is non-invasive, costs less, and fits better into routine clinical workflows Impellizzeri et al. (2020). However, compared to CBCT-based approaches, automated tooth-level analysis from IOS data remains underexplored, particularly for tooth presence classification, where limited annotated data and heterogeneous scan orientations pose substantial challenges.

The main unsolved challenge is automatically extracting a complete dental record from 3D scans. Specifically, a chart showing which teeth are present or absent. An automated system would give clinicians an immediate, accurate summary of a patient's dental status. This would reduce manual charting time, improve digital dentistry efficiency, and potentially reduce documentation errors. From a research perspective, this task also provides a challenging benchmark for studying learning under severe class imbalance and domain variability in medical 3D data.

1.2 Problem Statement

This project addresses the challenge of automating tooth presence classification from 3D intraoral scans. While human experts rely on visual inspection, automating this task using deep learning presents two fundamental technical conflicts that define our research problem:

1.2.1 The Representation Dilemma: Semantics vs. Geometry

The optimal data representation for dental automation remains an unresolved question in the medical AI community.

- **2D Render-based Approaches** leverage the well-established advantage of **transfer learning**. By projecting scans into 2D images, models can utilize powerful feature extractors (e.g., ResNet) initialized with weights pre-trained on large-scale external datasets like ImageNet. This allows the model to inherit robust visual feature extraction capabilities despite the limited size of medical datasets. However, this projection inevitably causes *information loss*, collapsing complex 3D spatial relationships into flat pixels.
- **3D Point Cloud Approaches** process the raw geometric data directly, preserving complete spatial fidelity. However, unlike 2D CNNs, specialized architectures like PointNet typically suffer from a **"cold start" problem**: they are often trained from scratch without the benefit of large-scale pre-trained foundation models. Consequently, they must learn both low-level geometric features and high-level semantic abstractions solely from the limited clinical dataset available.

The core research problem is therefore to determine whether the *intrinsic geometric integrity* of 3D processing outweighs the *feature-richness* provided by 2D transfer learning for the specific task of tooth presence detection.

1.2.2 Severe Class Imbalance in Clinical Data

Unlike standard classification tasks (e.g., cat vs. dog), tooth presence detection is a "needle in a haystack" problem. In a healthy population, missing teeth are statistically rare events.

- **Extreme Rarity:** In the 3DTeethSeg22 dataset, anterior teeth are missing in less than 5% of cases, while third molars are missing in over 30%.
- **The Accuracy Paradox:** Standard models trained on such distributions tend to collapse into trivial solutions, predicting "Present" for all teeth. This yields high accuracy (>95%) but fails at the clinical objective: identifying the few missing teeth.

Solving this requires not just a better model, but a fundamental rethinking of how loss functions and data synthesis can force networks to learn from minority examples.

1.3 Research Approach

Instead of an ad-hoc exploration, our investigation follows a systematic experimental design with controlled variables. We structured the research into four distinct phases to rigorously isolate the contributions of data representation, loss formulation, and data synthesis.

1.3.1 Phase 1: Baseline Representation Comparison

Goal: To explore baseline representation choices and identify structural limitations prior to focused model optimization.

- **Initial Baseline:** A standard 32-neuron output representation was implemented and evaluated as a diagnostic baseline to assess its suitability for single-jaw inputs.

- **2D Pathway:** ResNet-18 backbone initialized with ImageNet pre-trained weights, trained with standard Binary Cross-Entropy (BCE) loss on rendered images.
- **3D Pathway:** PointNet architecture trained from scratch, utilizing standard BCE loss on raw point clouds.
- *Objective:* This phase investigates whether the feature-richness of 2D transfer learning can plausibly compensate for the intrinsic geometric information loss incurred during projection.

1.3.2 Phase 2: Loss Function Optimization

We address the severe class imbalance by replacing the standard loss functions(BCE) with imbalance-aware alternatives(Dynamit Loss), keeping the dataset constant.

1.3.3 Phase 3: Data Augmentation

We introduce domain-specific synthetic data to explicitly increase the prevalence of minority classes (missing teeth).

- **Method:** A geometry-based "delete-and-fill" pipeline is applied to 3D meshes, in which selected teeth are removed and the resulting mesh cavities are sealed before rendering to 2D or sampling to point clouds.
- **Strategies:**

1. *Weighted Random:* Teeth are probabilistically removed from complete scans by randomly selecting K teeth (K = 2–5) using a weighted prior, with lower removal probabilities for wisdom teeth and third molars and higher probabilities for other teeth, aiming to mitigate severe class imbalance during training.

1. *Test-Informed Augmentation:* Missing-tooth patterns are sampled with guidance from the test label, providing clinically plausible combinations of absent teeth. These patterns are applied to training meshes via the same delete-and-fill process.

1.3.4 Phase 4: Exploration of Output Representations

Instead of fixing a single architecture, we investigated three different output configurations to assess the optimal representation for single-view tooth absence detection.

- **Baseline (32 Neurons):** Our initial approach employed a global 32-neuron output representing the full dentition, encoded as 0 for present and 1 for missing. However, since each input captures only one jaw, teeth from the opposing jaw are not visible in the input and are therefore encoded as 1, confounding structural absence with clinical tooth loss.
- **Jaw-Aware Extension (16+1 Neurons):** To improve upon the 32-neuron baseline, we extended the architecture to **16+1 neurons**. The added neuron acts as a jaw-type classifier (Upper vs. Lower). This design explicitly forces the network to recognize the anatomical orientation of the input scan, serving as a global context signal to assist the local tooth classification.

1.4 Scope and Limitations

This study addresses the binary classification of the presence of teeth for the 32 permanent adult teeth in full-arch scans. We do not evaluate tooth conditions, perform segmentation, or handle deciduous teeth.

The main limitation is the diversity of training data. Public IOS data sets do not represent the full spectrum of anatomical variations and clinical conditions. We partially addressed this through external validation on different scanners, but comprehensive generalization testing remains a topic of future work. Our 3D approach also requires pre-segmented meshes, limiting direct application to raw scans.

The study faces two primary data-related limitations. The dataset size is relatively restricted (900 scans), which may limit the model's exposure to rare anatomical variations found in a broader population. However, a more critical issue is the severe lack of positive samples. In the training set, missing teeth are extremely rare outside of wisdom teeth. This extreme imbalance creates a natural bias: the model learns that simply predicting 'present' is almost always correct, making it difficult to train effective detection for tooth loss without external intervention.

Chapter 2

Related Work and Research Gap

Automating dental analysis from digital scans represents an important next step in the evolution of digital dentistry. Although there has been considerable progress in segmenting and labeling teeth, the specific challenge of tooth *presence classification*, deciding which teeth are actually present or missing, has received little direct attention. In this chapter, we review current research in dental AI, highlight important methodological discussions in 3D medical data processing, examine how researchers tackle severe class imbalances, and clearly state the remaining research gaps that our work addresses.

2.1 Deep Learning in Digital Dentistry

2.1.1 Automated Tooth Analysis from 3D Scans

Recent advances in deep learning have enabled automated analysis of 3D dental models, particularly in the domain of tooth segmentation and labeling. The majority of existing research treats tooth analysis as a *geometric partitioning problem*, where the primary objective is to delineate individual tooth boundaries and assign anatomical labels to visible structures.

Dominance of Segmentation Tasks

The 3DTeethSeg'22 challenge Ben-Hamadou et al. (2023a) exemplifies the current research focus. It is centered exclusively on segmentation and labeling of *existing teeth* from intraoral scan (IOS) data, implicitly formulating the problem only over visible anatomical structures, without an explicit representation for tooth absence. Participating methods achieved impressive performance on this task, with state-of-the-art approaches reaching high Dice coefficients Chen et al. (2023).

Building on mesh-based deep learning, Lian et al. Lian et al. (2020) introduced MeshSegNet, a pioneering graph neural network (GNN) architecture that processes raw 3D dental meshes directly. Their method integrates graph-constrained learning modules to hierarchically extract multi-scale local contextual features, achieving strong segmentation performance on intraoral scans. However, we note that MeshSegNet suffers from heavy computational requirements due to large-scale matrix operations on adjacency matrices. Subsequent refinements by Wu et al. Wu et al. (2021) proposed a two-stage framework (TS-MDL) for joint tooth labeling and landmark localization, demonstrating that mesh deep learning can extend beyond segmentation to anatomical feature detection.

More recent approaches have explored advanced architectures to handle the complexity of dental geometry. Ghamrawi et al. Ghamrawi et al. (2024) introduced TSegLab, a multi-stage framework that further refines the segmentation and labeling process. Additionally, Zhu et al. Zhu et al. (2026) proposed a high-fidelity reconstruction method fusing IOS and CBCT data via deep implicit representations, indicating a shift towards more comprehensive 3D dental analysis.

Hierarchical self-supervised learning frameworks have also emerged to address the scarcity of labeled IOS data. Zhou et al. Zhou et al. (2023) introduced STSNet, which employs point-level, region-level, and cross-level contrastive losses for unsupervised representation learning. They demonstrated that pre-training on unlabeled data substantially improves segmentation performance with limited annotations.

Despite these advances, we identify a critical limitation that pervades the literature: **the majority of existing methods assume the presence of teeth to be segmented**. They focus on delineating boundaries of *visible* anatomical structures but do not address the clinical scenario where teeth are *absent*. As Xu et al. Xu et al. (2021) noted in their clinical validation study, segmentation models perform well under normal conditions but exhibit degraded accuracy when confronted with abnormal dentition patterns, including missing teeth. We agree with this observation and argue that segmentation alone is insufficient for complete dental diagnosis.

The Absence Detection Gap

Clinically, dental charting requires not only identifying present teeth but also detecting tooth absence, which is a fundamentally different task. Missing teeth leave no explicit tooth geometry in the scan and must be inferred indirectly from gaps, surface topology, and anatomical context. We believe this inference problem differs fundamentally from segmentation in three key ways:

- **Negative evidence**: Segmentation operates on positive visual features such as crown morphology and surface curvature. In contrast, absence detection requires recognizing the *lack* of expected structures, a form of negative evidence that standard segmentation architectures are not designed to capture.

- **Contextual reasoning**: Determining whether a gap represents a missing tooth versus normal spacing requires an understanding of dental anatomy and typical tooth arrangements. We observe that segmentation models trained on pixel or vertex labels do not explicitly learn this contextual information.

- **Class imbalance**: In population-level data, missing teeth (particularly anterior teeth) are statistically rare events, often with less than 5% prevalence in healthy populations. This creates a severe training imbalance that segmentation objectives do not adequately address.

While some studies tangentially address missing teeth, they do so as an auxiliary byproduct rather than a primary objective. For instance, recent work on tooth numbering systems Chung et al. (2021) incorporates missing tooth detection as a post-processing step to correct numbering inconsistencies. However, we contend that relying on heuristic rules rather than learned detection models limits the robustness of these approaches.

To our knowledge, existing IOS-based studies do not treat tooth presence classification as a standalone learning task. The few existing studies on missing tooth detection operate exclusively in the 2D radiographic domain, which we discuss next.

2.1.2 2D Radiographic Approaches: Insights and Limitations

The detection of missing teeth has received more attention in 2D dental radiography, providing valuable methodological insights despite operating on fundamentally different data modalities.

X-ray Based Methods

Several recent studies have applied deep learning to detect missing teeth in panoramic radiographs (2D X-ray images). Park et al. Park et al. (2022) proposed a two-stage pipeline for dental implant planning. First, a Mask R-CNN model segments individual teeth, and second, a detection module identifies regions of missing teeth based on gaps in the segmentation output. Their method achieved 92.14% mean average precision (mAP) for tooth segmentation but only 59.09% mAP for missing tooth region detection. We interpret this discrepancy as clear evidence that absence detection remains substantially more challenging than presence-based segmentation.

Kim et al. Kim et al. (2024) focused specifically on pediatric patients, where early tooth loss is common. They developed a quadrant-based classification system that predicts whether each of four jaw regions contains missing teeth. Using DenseNet169 on panoramic radiographs, they achieved a classification accuracy of 73% and an F1-score of 0.731. Notably, their analysis revealed region-specific performance variations: anterior teeth (canines, premolars) showed higher detection accuracy than posterior molars. They attributed this to clearer visual boundaries in radiographic images, a finding we find consistent with general radiographic properties.

Al-Sarem et al. Al-Sarem et al. (2022) conducted a systematic comparison of six pretrained CNN architectures (AlexNet, VGG16, VGG19, ResNet50, DenseNet169, MobileNetV3) for classifying missing tooth regions in CBCT-derived 2D slices. Their best model (DenseNet169) achieved 93.3% segmentation accuracy and 89% classification accuracy for missing tooth regions, significantly outperforming shallow architectures. We take this work as a strong indication of the value of transfer learning from ImageNet-pretrained models, even in the specialized domain of dental imaging.

Methodological Lessons from 2D Studies

These 2D radiographic studies provide three key insights relevant to our IOS-based investigation:

1. **Absence detection is harder than presence detection**: Across all studies, performance metrics for missing tooth detection consistently lag behind those for tooth segmentation by 10 to 30 percentage points. We believe this indicates that recognizing negative evidence requires fundamentally different learning strategies.

2. **Transfer learning provides substantial benefits**: Pretrained CNN backbones (e.g., ResNet, DenseNet) initialized with ImageNet weights consistently outperform models trained from scratch. This demonstrates that general-purpose visual features learned from natural images can transfer effectively to dental imaging.

3. **Class imbalance is a pervasive challenge**: Studies explicitly acknowledge the severe imbalance between present and missing teeth, employing strategies such as weighted loss functions and oversampling. However, we note that most rely on relatively balanced datasets (30 to 50% positive samples), avoiding the extreme scarcity (5% or less) encountered in real-world anterior tooth absence scenarios.

The IOS vs. Radiography Gap

Despite these methodological contributions, we argue that 2D radiographic approaches cannot be directly applied to IOS data due to fundamental modality differences:

- **Dimensionality**: Radiographs are 2D projections capturing sub-gingival anatomy (roots, alveolar bone), whereas IOS provides high-resolution 3D surface geometry of supra-gingival structures (crowns, occlusal surfaces). Missing tooth detection from IOS relies on surface topology and spatial arrangement rather than radiographic density patterns.
- **Clinical workflow integration**: IOS is routinely performed in general dental practice for restorative and orthodontic procedures, whereas CBCT and panoramic radiographs require specialized imaging equipment and are typically reserved for surgical planning. We believe an IOS-based detection method would integrate more seamlessly into existing digital dentistry workflows.
- **Occlusion and incomplete coverage**: Each IOS scan captures only one jaw (maxillary or mandibular), requiring models to distinguish between structural absence (opposing jaw not visible) and clinical tooth loss. This introduces a unique challenge absent in full-arch radiographic images.

Critical Gap: While radiographic studies demonstrate the feasibility of learning-based missing tooth detection, the task remains largely unexplored for IOS data. We maintain that the different geometric representation (3D surface meshes vs. 2D images) and unique data characteristics (single-jaw coverage, high-resolution surface detail) necessitate a dedicated investigation.

2.1.3 Research Gap Summary

Having established that (i) 3D tooth segmentation methods ignore absence detection and (ii) 2D radiographic methods cannot transfer to IOS, we explicitly define the core research gaps that motivate this project.

Unlike segmentation tasks that produce dense labels based on positive visual features, **Tooth Presence Classification** requires recognizing *negative evidence* (the absence of expected structures) and producing a sparse binary vector over predefined tooth positions. This task is clinically significant for automating dental charting but faces unique challenges that existing literature has not addressed:

1. **Lack of dedicated IOS presence classifiers**: No prior work has treated tooth presence as a primary classification task on 3D surface scans.
2. **Generalization failure on rare patterns**: As hinted by 2D studies, standard models struggle with the extreme class imbalance inherent in tooth loss patterns (e.g., rare anterior missing teeth).
3. **Methodological uncertainty**: It remains unclear whether 3D data should be processed in its native form or projected into 2D to leverage pretrained networks for this specific task.

2.1.4 Summary and Transition

This section establishes that while 3D tooth segmentation from IOS has matured significantly Ben-Hamadou et al. (2023a); Lian et al. (2020); Wu et al. (2021), the task of determining tooth presence

remains largely unexplored in the IOS domain. 2D radiographic studies Park et al. (2022); Kim et al. (2024); Al-Sarem et al. (2022) demonstrate the feasibility of learning-based absence detection but operate on fundamentally different data modalities.

The following sections will address the identified gaps by examining two critical methodological questions: (i) whether to process IOS data in its native 3D form or project it into 2D for leveraging pretrained models, and (ii) how to effectively address the severe class imbalance inherent in missing-tooth detection.

Chapter 3

Methodology

This chapter presents the complete methodology for both the 2D render-based and 3D point cloud-based pathways for automated tooth presence classification, with a particular focus on missing-tooth detection under severe class imbalance. We first describe the datasets and preprocessing procedures, including the specialized data augmentation pipeline developed to address class imbalance. We then detail the technical approach of each pathway, and finally present the training and optimization strategies.

3.1 Datasets and Data Preprocessing

3.1.1 Label Preprocessing and Output Evolution

The raw segmentation data is converted into binary vectors where $y = 1$ indicates a missing tooth (positive class) and $y = 0$ indicates a present tooth (negative class). Our output configuration evolved through three stages to handle the single-view nature of the input images:

- **Global Representation (32 neurons):** Initially, we employed a standard 32-dimensional output vector representing the full dentition (FDI 11–48). However, since each 2D input image captures only a single dental arch (maxillary or mandibular), half of the label space was systematically unobservable, leading to optimization redundancy.
- **Jaw-Aware Representation (16+1 neurons):** Our second architecture, utilized in the reported experiments, extends the output to 17 neurons. The first 16 neurons predict tooth absence for the visible arch, while the 17th neuron serves as a binary classifier for the jaw type (0 for upper, 1 for lower). This architecture is trained in different experimental settings, including binary cross-entropy (BCE), Dynamit-based loss reweighting for tooth output, and Dynamit-based training with augmented synthetic data.

3.1.2 Training Dataset

Our primary training dataset is the **3DTeethSeg22 challenge dataset** Ben-Hamadou et al. (2023b)Ben-Hamadou et al. (2022), which comprises **1,800 intraoral scans from 900 patients** (specifically, 900 upper jaw and 900 lower jaw scans). Each scan is provided as a triangular mesh in the `.obj` format, with corresponding per-vertex tooth labels given in the `.json` format. These labels follow the **FDI (Fédération Dentaire Internationale) two-digit notation system**. **In this system, the oral cavity is divided into four quadrants, where the first digit denotes the quadrant (1-4 for permanent teeth) and the second digit identifies the specific tooth (1-8). Figure 3.1 illustrates**

the adult FDI numbering convention.

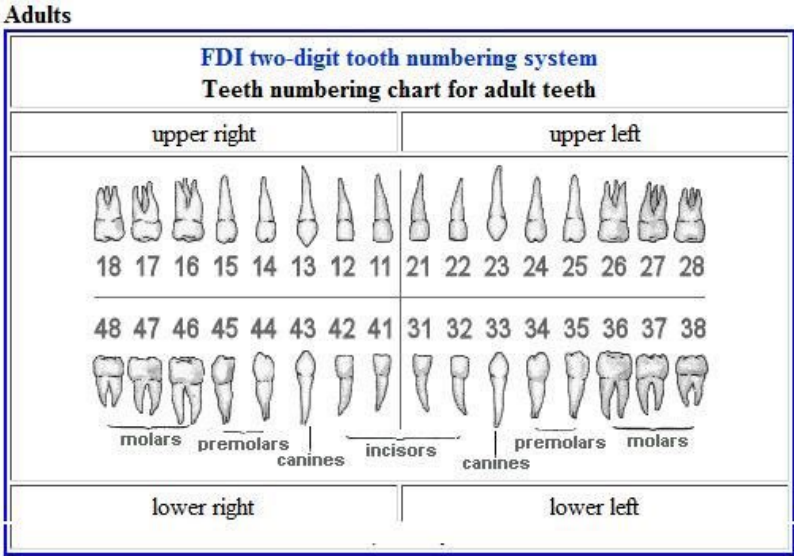


Figure 3.1: **The international tooth numbering system (FDI)** (Truant (nd))

To establish the optimal learning target, we analyzed the training set label distribution under two potential encoding schemes: a global 32-neuron representation covering the full dentition, and a jaw-specific 17-neuron representation comprising 16 teeth and a jaw indicator.

Due to the single-jaw nature of the dataset, each scan captures only one jaw, teeth from the opposing arch are structurally absent. In a global 32-neuron encoding, these out-of-field teeth are assigned the "Missing" label, artificially inflating the "Missing" class. For example, Tooth 31 is clinically missing in only 0.33% of lower jaw cases (3 out of 900). However, a global encoding would incorrectly label it as "missing" in all upper jaw scans, inflating its rate to $\approx 50\%$.

To avoid conflating structural absence with clinical pathology, we adopted the 17-neuron jaw-specific configuration (16 teeth + jaw) as the target label encoding. Having defined the jaw-specific encoding, we visualized the training set distribution ($N = 1800$).

Figure3.2 illustrates the contrast between naive global counting and jaw-specific counting, demonstrating the correction of artificially inflated absence labels.

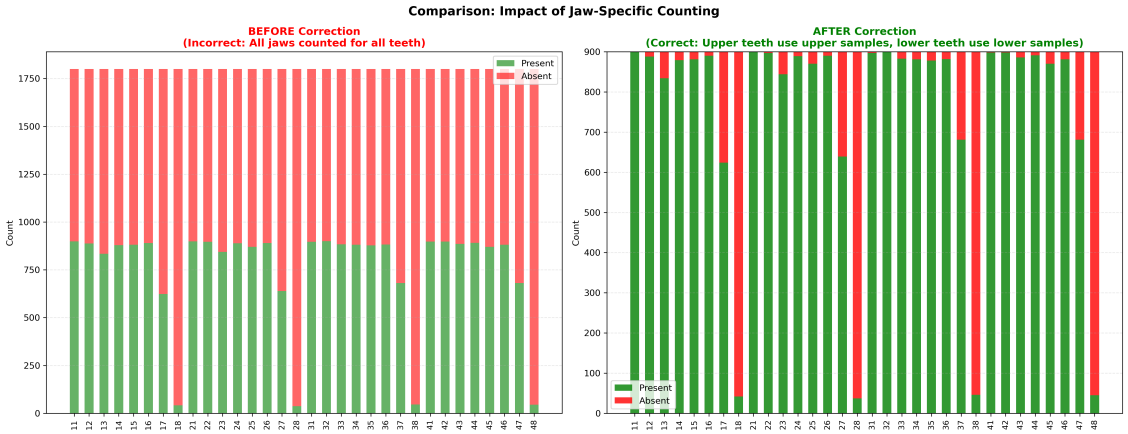


Figure 3.2: **Comparison of tooth presence and absence counts before and after applying jaw-specific counting. Green bars** indicate present teeth, while **red bars** indicate absent teeth. **Left**: Naive global counting, where all 32 teeth are evaluated for every sample regardless of jaw visibility. Teeth from the opposing jaw are incorrectly labeled as absent, leading to systematic inflation of missing counts. **Right**: Corrected jaw-specific counting, where upper-jaw teeth are computed exclusively from upper-jaw samples and lower-jaw teeth from lower-jaw samples.

Figure 3.3 visualizes the jaw-separated presence–absence distribution, highlighting the strong anatomical and statistical asymmetry between common teeth and wisdom teeth.

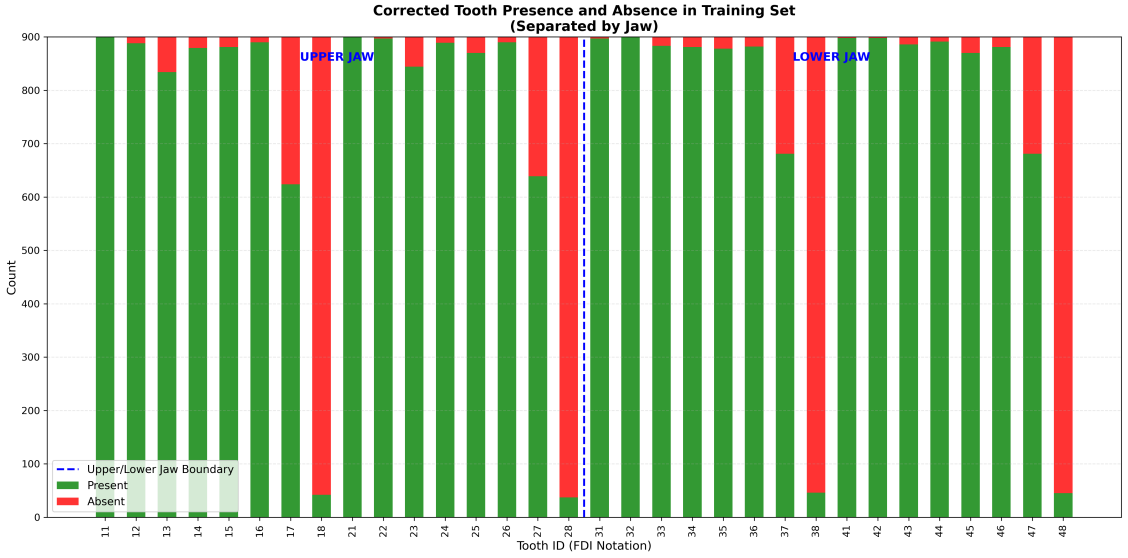


Figure 3.3: **Stacked distribution of tooth presence and absence in the corrected training set, separated by jaw. Green segments** indicate present teeth, and **red segments** indicate absent teeth. Upper-jaw teeth (FDI 11–28) and lower-jaw teeth (FDI 31–48) are separated by the dashed vertical line.

Overall, the corrected training set exhibits a highly imbalanced label distribution, with the vast majority of teeth being present and missing labels occurring only in a small subset of teeth.

3.1.3 Testing Dataset

For external validation and generalization assessment, we use a private clinical dataset of approximately 180 intraoral scans. This authorized data originates from the University Hospital Zürich (USZ). The Real World clinical data were acquired using different scanner hardware and originate from a different clinical source than the training data, making it an ideal and robust benchmark for evaluating model performance on data from a different clinical environment.

The test set underwent the same cleaning and preprocessing procedure as the training set. After matching rendered images with CSV entries and removing duplicate samples, the final test set consists of 167 unique samples, including 87 upper-jaw scans and 80 lower-jaw scans.

To ensure consistency with the training setup, jaw-specific counting was applied when computing tooth-level statistics. Specifically, upper-jaw teeth were evaluated only in upper-jaw samples, and lower-jaw teeth were evaluated only in lower-jaw samples. This prevents structurally unobserved teeth from being incorrectly treated as clinically absent.

To illustrate the effect of jaw-specific counting on test-set statistics, we visualize the tooth-level distributions before and after applying the correction. As shown in Figure 3.4, naive counting that evaluates all teeth in all samples leads to an inflated number of absent labels, while jaw-specific counting produces a distribution that reflects anatomical observability.

Figure 3.4: **Comparison of tooth presence and absence counts in the test set before and after applying jaw-specific counting. Green bars** denote present teeth, and **red bars** denote absent teeth. **Left**: Naive counting where all samples are used for every tooth, regardless of jaw visibility. **Right**: Jaw-specific counting where teeth are evaluated only within their anatomically observable jaw.

Figure 3.5 presents the jaw-separated tooth presence and absence distribution in the corrected test set. The bar chart shows the number of samples in which each tooth is clinically absent for each FDI position. Consistent with clinical prevalence, third molars exhibit the highest absence

frequency, while other tooth positions maintain a relatively stable representation across samples.

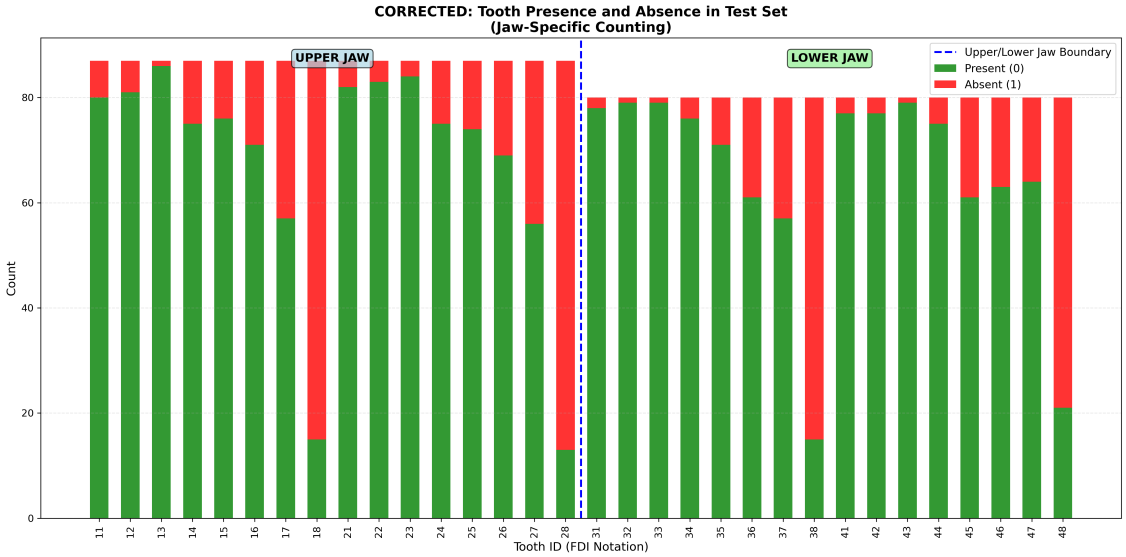


Figure 3.5: **Stacked distribution of tooth presence and absence in the corrected test set, separated by jaw. Green segments** indicate present teeth, and **red segments** indicate absent teeth. Upper-jaw and lower-jaw teeth are divided by the dashed vertical line.

3.1.4 Dataset with Data Augmentation

To mitigate the severe class imbalance observed in the original training data, we construct augmented datasets using two different augmentation strategies. Both strategies aim to increase the representation of missing teeth while preserving anatomical plausibility.

The first strategy, referred to as test-pattern-based augmentation, follows clinically observed tooth absence patterns derived from the test set. The second strategy applies random tooth removal with predefined probabilities per tooth type. In both cases, augmentation is performed at the label level, and jaw-specific counting is consistently applied to ensure that only anatomically observable teeth are considered.

We first visualize the overall tooth presence and absence distribution after augmentation for both strategies. Both augmentation strategies substantially increase the number of missing-tooth samples compared to the original training set. To further examine how the two augmentation strategies affect individual tooth distributions, we compare the per-tooth missing rates.

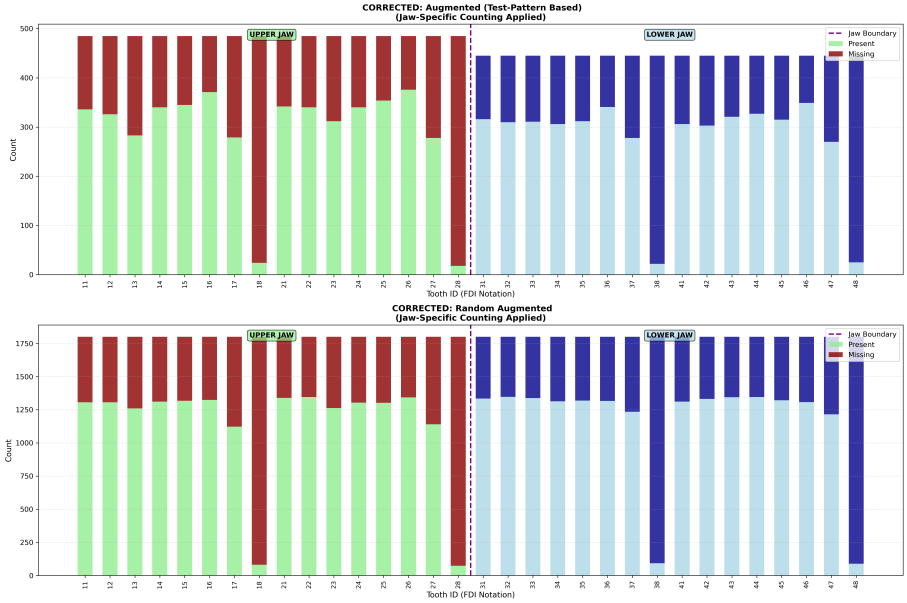


Figure 3.6: **Tooth presence and absence distributions after augmentation using two strategies with jaw-specific counting applied. Top**: Test-pattern-based augmented dataset. **Bottom**: Randomly augmented dataset. Upper-jaw and lower-jaw teeth are separated by the dashed vertical line.

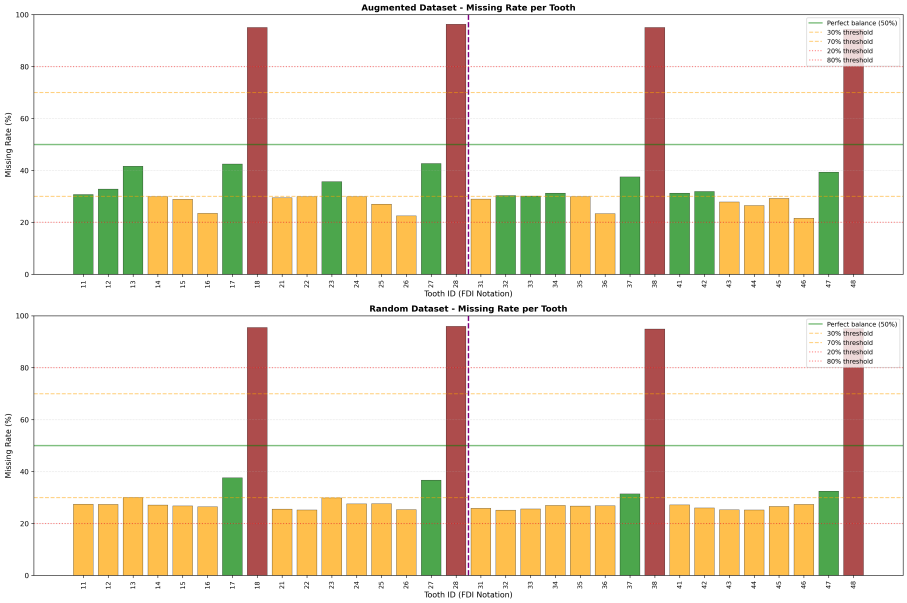


Figure 3.7: **Per-tooth missing rate comparison between test-pattern-based augmentation and random augmentation.** Bars indicate the percentage of missing samples for each tooth under the two augmentation strategies.

While both strategies increase missing rates across most teeth, the resulting distributions differ a bit in their tooth-specific balance profiles.

Data augmentation substantially alters the label distribution and reduces the extreme imbalance present in the original training set. However, a uniformly balanced distribution across all teeth cannot be achieved, as posterior teeth are inherently more likely to be absent due to anatomical and clinical factors.

3.2 3D Point Cloud-Based Classification Pathway

The 3D pathway processes dental scan data directly in its native geometric representation as unordered point clouds. This approach is preferred as it leverages architectures specifically designed for 3D point set learning, thereby **preserving intrinsic geometric details** and avoiding information loss inherent in intermediate representations such as voxelization or multi-view projection.

3.2.1 Point Cloud Preprocessing

Raw point clouds extracted from intraoral scans exhibit significant variations in density, scale, and orientation. To ensure geometric consistency and facilitate feature learning, we implement a standardized preprocessing pipeline consisting of three steps:

1. Canonical Orientation via PCA Since PointNet is sensitive to rigid transformations, we use Principal Component Analysis (PCA) to align each point cloud's principal axes with the coordinate frame, ensuring the dental arch curve aligns with the X-Y plane and tooth height aligns with the Z-axis.

2. Jaw-Specific Normalization Point clouds are centered at the origin and scaled to a unit sphere by normalizing by the maximum distance from the centroid. By Normalization the geometric pose of both jaws, we allow the network to learn shared feature representations for tooth morphology, significantly reducing the complexity of the classification task.

3. Uniform Sampling Finally, to ensure scale and translation invariance, all point clouds are centered at the origin and scaled to fit within a unit sphere. The data is then uniformly resampled to a fixed size of $M = 4096$ points. This resolution was empirically chosen to balance computational efficiency with the preservation of fine geometric details required for distinguishing adjacent teeth.

Network Architecture

We employ a PointNet-based architecture Qi et al. (2016) designed to process unordered 3D point sets while maintaining permutation invariance. The network consists of two primary components: a feature encoder and a classification head.

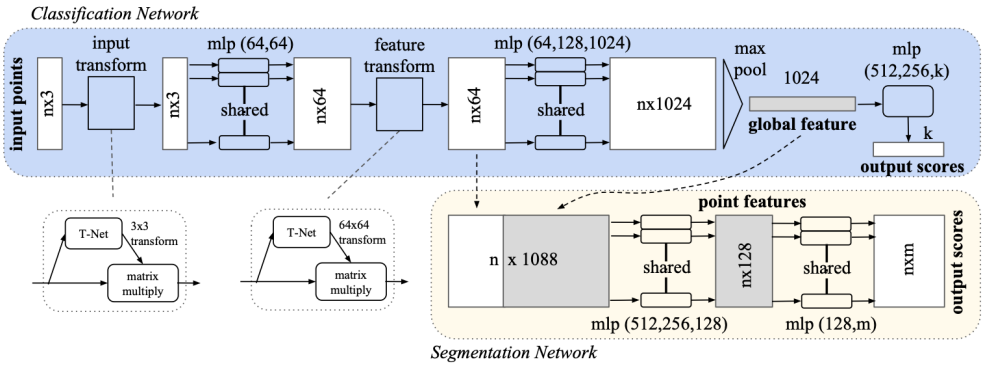


Figure 3.8: **PointNet Architecture.** The classification network takes n points as input, applies input and feature transformations, and then aggregates point features by max pooling. The output is classification scores for k classes. The segmentation network is an extension of the classification net. It concatenates global and local features and outputs per point scores. "mlp" stands for multi-layer perceptron, numbers in brackets are layer sizes. Batchnorm is used for all layers with ReLU. Dropout layers are used for the last mlp in the classification net. (Image from Qi et al. (2016))

PointNet Encoder The encoder maps the input point cloud $P = \{p_i \in \mathbb{R}^3\}_{i=1}^N$ into a global feature vector. To respect the permutation invariance of the input set, the network approximates a general function f defined on a point set as:

$$f(\{p_1, \ldots, p_n\}) \approx g(h(p_1), \ldots, h(p_n)) \tag{3.1}$$

where h is a spatial encoding function and g is a symmetric aggregation function.

In our implementation, h is realized by a series of shared Multi-Layer Perceptrons (MLPs) that process each point independently. This lifts the individual point coordinates into a high-dimensional feature space (typically up to 1024 dimensions). Subsequently, a max-pooling operation serves as the symmetric function g:

$$\mathbf{g} = \max_{i=1,\ldots,N} h(p_i) \in \mathbb{R}^{1024}$$

This operation aggregates point-wise features into a single global descriptor $\mathbf{g}$, ensuring the representation remains unchanged regardless of the input point ordering.

Classification Head The global feature vector $\mathbf{g}$ is then processed by a classification sub-network. Following the original architecture design, this consists of fully connected (FC) layers with progressively decreasing dimensions ($1024 \to 512 \to 256 \to k$).

To adapt the model for our tooth presence task, the final layer is configured to output $k = 32$ logits, corresponding to the 32 permanent tooth positions. Batch normalization is applied to all layers with ReLU activations to stabilize training, and dropout (rate=0.3) is employed in the final fully connected layers to prevent overfitting.

3D Tooth Absence Simulation and Data Augmentation

To mitigate severe class imbalance, we implemented a geometry-based augmentation strategy operating directly on 3D meshes. By surgically removing teeth and sealing the resulting geometry, we generate anatomically plausible missing tooth samples prior to point cloud conversion.

The "Remove-and-Fill" Pipeline We devised a unified mesh processing pipeline to simulate tooth extraction. To intuitively demonstrate the quality of this augmentation, we present a rendered visualization of a processed mesh in Figure 3.9.

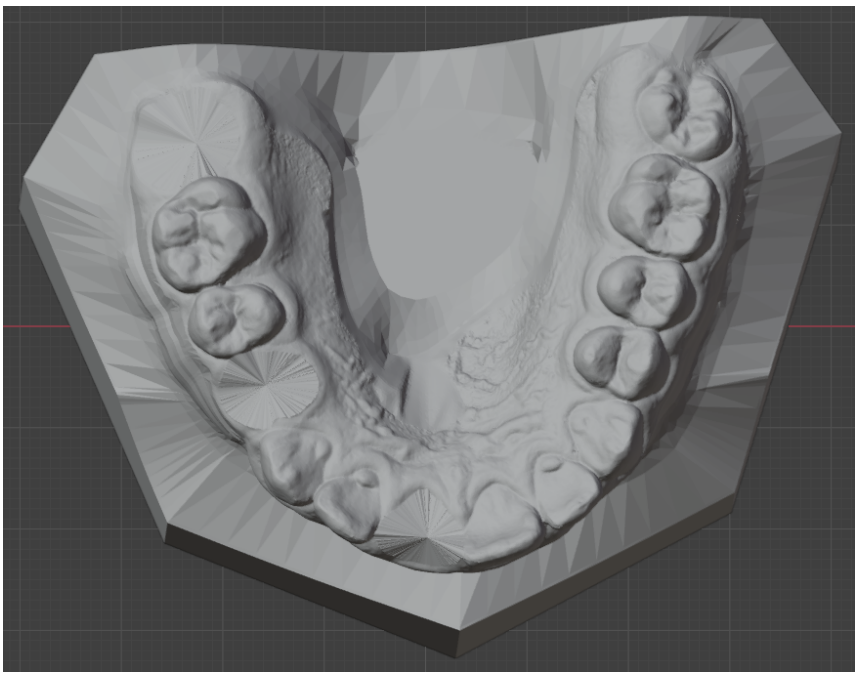


Figure 3.9: **Visualization of the 3D augmentation result.** This rendered view intuitively demonstrates the efficacy of the "Remove-and-Fill" pipeline. The target tooth has been removed, and the socket is visually sealed to obtain an anatomically plausible representation of tooth absence.

The pipeline consists of three steps:

1. **Extraction:** Vertices belonging to the target tooth are identified via labels and removed.

2. **Sealing:** Boundary edges of each connected boundary component are extracted. A centroid vertex is introduced, and cap triangles are generated per boundary edge. The orientation of newly created faces is corrected using normals of adjacent retained faces, with a PCA-based fallback to ensure consistent face orientation and reduce visual holes caused by backface culling.

3. **Healing:** The newly created faces are assigned the gingiva material, and vertices associated with removed teeth are labeled as gingiva (0), yielding a visually healed socket with reduced geometric artifacts.

Dual Augmentation Strategy We employed two complementary strategies to balance randomization with clinical realism. Table 3.1 details the generation logic and yield for each method.

Table 3.1: Comparison of Data Augmentation Strategies

Strategy	Generation Logic	Parameters	Yield
Probability-Driven	**Goal:** Diversity via randomization. Randomly removes 2–5 teeth based on clinical frequency weights (weighted probabilistic removal with tiered probabilities: wisdom teeth $P = 0.02$, second molars $P = 0.12$, and other teeth base $P = 0.30$). For each augmented sample, we sample K in [2,5] and select teeth via weighted sampling. Due to duplicate draws, the number of unique removed teeth may occasionally be smaller than K.	2 copies per scan	3,600
Pattern-Driven	**Goal:** Clinical realism. Replicates 186 missing-tooth patterns derived from the held-out test-set label distribution.	5 copies per pattern	930

Final Training Set Composition The final dataset combines the original scans with both augmented subsets:

$$\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{original}}(1,800) \cup \mathcal{D}_{\text{random}}(3,600) \cup \mathcal{D}_{\text{pattern}}(930)$$

This results in a total of **6,330 samples**, significantly enriching the representation of the minority (missing) class while preserving anatomical validity.

3.3 2D Render-Based Classification Pathway

The 2D pathway transforms the 3D tooth presence classification problem into a multi-label 2D image classification task. This approach leverages the success of convolutional neural networks (CNNs) in image recognition while benefiting from pre-training on large-scale datasets such as ImageNet. The key innovation is a high-fidelity Blender rendering pipeline that converts 3D meshes into diagnostically informative 2D renderings under a standardized canonical view, with model-specific test-set rotations to handle unconstrained scan orientations.

3.3.1 Motivation for 2D Representation

While 3D point clouds preserve complete geometric information (Section 3.2), 2D image-based approaches offer several complementary advantages:

- **Transfer learning:** Pre-trained CNN models (e.g., ResNet, EfficientNet) trained on millions of natural images provide strong low-level feature extractors that can be fine-tuned for dental imaging.
- **Computational efficiency:** 2D convolutions are more mature and optimized in deep learning frameworks compared to 3D point cloud processing.
- **Visual interpretability:** 2D renderings facilitate human interpretation and validation of model predictions, supporting qualitative inspection and potential clinical use.

Our 2D pathway consists of three key components: a rendering pipeline to generate multi-view 2D representations, a CNN-based classification architecture, and domain-specific data augmentation strategies.

3.3.2 3D-to-2D Rendering Pipeline

We developed a Blender-based rendering pipeline to convert 3D dental mesh models into high-quality 2D images suitable for training deep networks. The pipeline is designed to preserve diagnostically relevant geometric features while producing visually consistent and high-fidelity images across the entire dataset. A separate five-view pilot rendering was first conducted for visual inspection and view selection, after which the final data set for training generation was standardized to a single canonical view.

Pose Normalization

In practice, we first apply a lightweight heuristic pre-alignment based on mesh extents and surface normal statistics to obtain a stable initial top-view orientation. To ensure consistent orientation across all 3D samples entering the rendering pipeline, we apply the PCA-based pose normalization, as detailed in Section 3.2.1:

1. A stable local frame is derived by performing PCA on the mesh vertices.
2. Construct an orthonormal basis by aligning the Z-axis with the direction of least variance. The X and Y axes are then determined from the projected extents of the remaining eigenvectors, ensuring a right-handed system.
3. Apply an additional mandibular-specific flip check to ensure the occlusal surface faces +Z.

This preprocessing step ensures that the rendered top views correspond to comparable anatomical orientations across samples, which is essential for using a single fixed camera configuration in the subsequent 2D rendering stages.

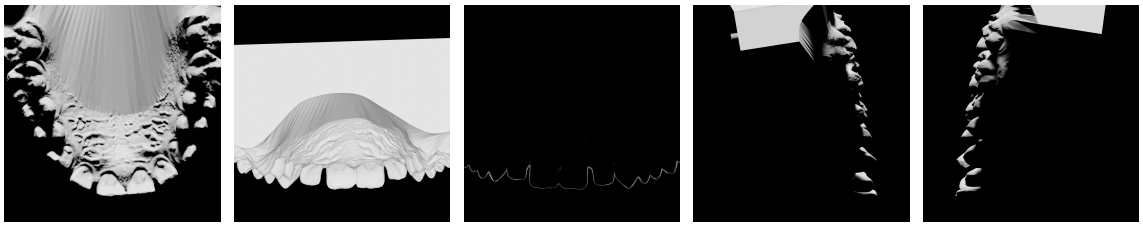
View Selection via Multi-View Rendering

Prior to final dataset generation, we conducted a separate multi-view rendering experiment only for viewpoint comparison and selection for training set. In this preliminary analysis, each 3D mesh model can be rendered from five orthographic viewpoints defined by camera position along the canonical axes:

1. **Superior view** (top): Orthographic view from above the dental arch.
2. **Inferior view** (bottom): Orthographic view from below the dental arch.
3. **Anterior view** (front): Orthographic view along the anterior direction of the dental arch.
4. **Left lateral view**: Orthographic view exposing the left side of the dental arch.
5. **Right lateral view**: Orthographic view exposing the right side of the dental arch.

Scale consistency:
Orthographic projection maintains constant scale across all regions of the dental arch, ensuring that distal teeth (molars) and anterior teeth (incisors) occupy proportional image areas. This is crucial for uniform feature extraction by the CNN.

Metric preservation:
Unlike perspective projection, which introduces foreshortening effects, orthographic projection preserves relative distances and angles, facilitating more consistent geometric feature learning. The camera distance and orthographic scale are configured to ensure the entire dental arch fits within the viewport with minimal background.



(a) Superior (Top) (b) Inferior (Bottom) (c) Anterior (Front) (d) Left Lateral (e) Right Lateral

Figure 3.10: **Comparison of five orthographic viewpoints.** The Superior (top) view in (a) provides the most informative and consistent representation of tooth morphology and tooth presence cues. In contrast, the Anterior (c) and Lateral views (d, e) suffer from severe occlusion, while the Inferior view (b) mainly exposes the surface opposite to the occlusal plane and is less informative for tooth presence identification.

This five-view in Figure 3.10 rendering was used exclusively for qualitative analysis and viewpoint selection, and is not used for final training or testing.

Training-Set Rendering (Top View):

We employ orthographic projection to preserve the geometric proportions of the dental structures. Initially, we implemented a five-view rendering setup (top, bottom, front, left, right) as a preliminary exploration step, hypothesizing that a multi-view approach might maximize feature extraction.

However, a visual analysis of these preliminary multi-view rendering images (Figure 3.10) revealed significant redundancy and a lack of information in most views. Specifically, the Inferior view (Figure 3.10b) captures the internal face of the mesh, resulting in a negligible signal. Similarly, the Anterior and Lateral views (Figure 3.10c, 3.10d, 3.10e) exhibit limited informational value due to self-occlusion of the teeth. Consequently, we found that the **Superior** (top) view, illustrated in Figure 3.10a, provides the most distinct and comprehensive representation of the dental features. Based on this observation, all subsequent training set was rendered exclusively from the **Superior** (top) view (Figure 3.11).

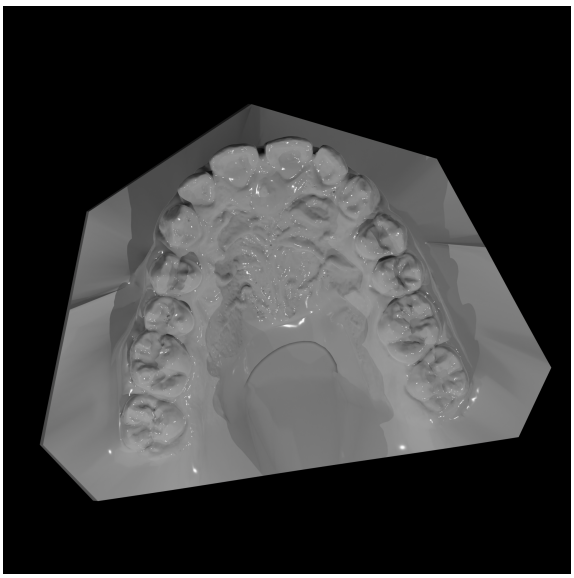


Figure 3.11: **Example of training-set 2D renderings**

After selecting the **Superior** (top) view, the same rendering pipeline, which uses identical camera configuration, lighting, and material settings is also applied to augmented training meshes, ensuring consistent appearance across all training samples (Figure: 3.12).

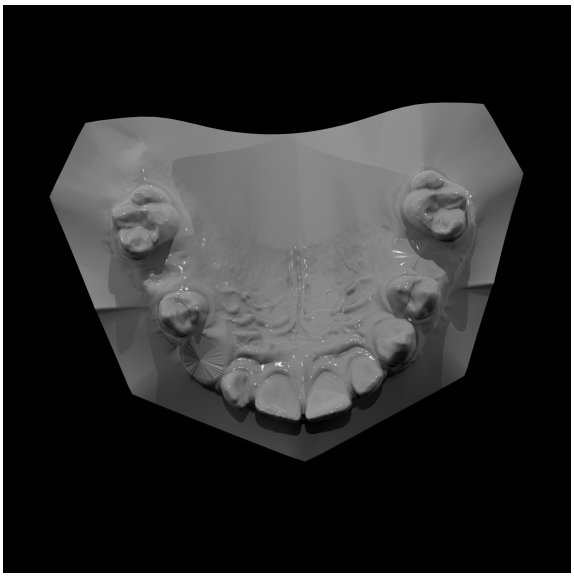
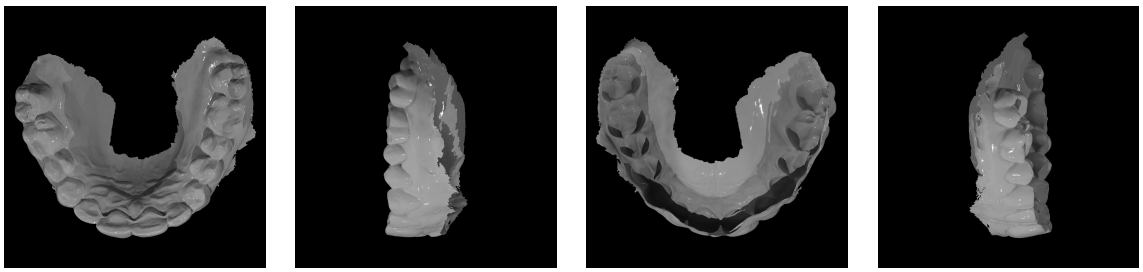


Figure 3.12: **Example of augmented training-set 2D renderings**

Test-set Rendering for 32-Neuron Model (Preliminary In-Plane Rotations):

Unlike the training and augmented scans, the 3D meshes in the test set do not have a standardized orientation. After projecting a 3D object into a 2D image, we cannot freely change the 3D viewpoint anymore. Therefore, to maximize tooth visibility and reduce failures caused by unfavorable orientations, we render the test set using the same superior (top) view but under four in-plane rotations around the vertical axis: $0°$ (Figure 3.13a), $90°$(Figure 3.13b), $180°$(Figure 3.13c), and $270°$(Figure 3.13d). This produces four candidate images per test scan and increases the chance that most teeth are visible in at least one rendering. This strategy is reported only to document the initial diagnostic baseline exploration and is not part of the final quantitative evaluation.



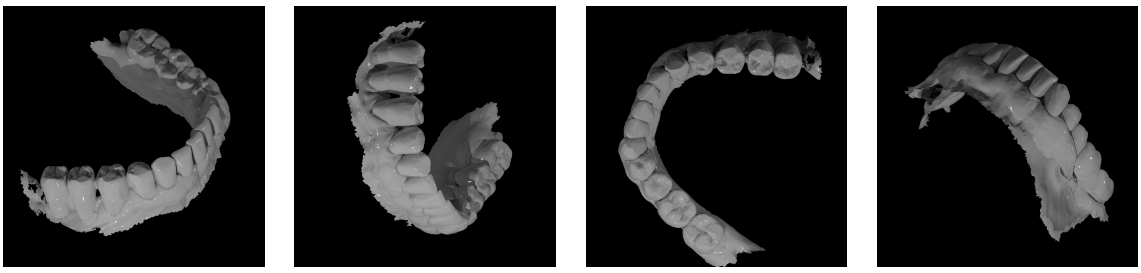
(a) $0°$ (b) $90°$ (c) $180°$ (d) $270°$

Figure 3.13: **Example of test-set rendering under four in-plane rotations**

Test-Set Rendering for 16+1 Jaw-Specific Model:

For the $16 + 1$ jaw-aware architecture, each input image corresponds to a single dental arch, making the model inherently sensitive to the pose and orientation of the underlying 3D scan. However, unlike the training data, the test-set meshes are not consistently aligned, which may cause a single fixed projection to fail in exposing sufficient anatomical cues in some cases.

To analyze this sensitivity, we first conducted a preliminary exploration using a predefined set of 24 discrete 3D rotations. This rotation set includes multiple in-plane rotations around the vertical axis as well as additional out-of-plane rotations along the x- and y-axes. The goal of this exploration was to qualitatively assess how different viewing angles affect the visibility of tooth morphology and global arch structure in 2D renderings. Representative examples of the resulting projections under different rotation configurations are shown in Figure 3.14.



(a) Tilted view obtained by rotating the scan by -90° around the y-axis

(b) Top view with no tilt applied

(c) Tilted view obtained by rotating the scan by $+90^\circ$ around the y-axis

(d) Tilted view obtained by rotating the scan by $+90^\circ$ around the x-axis

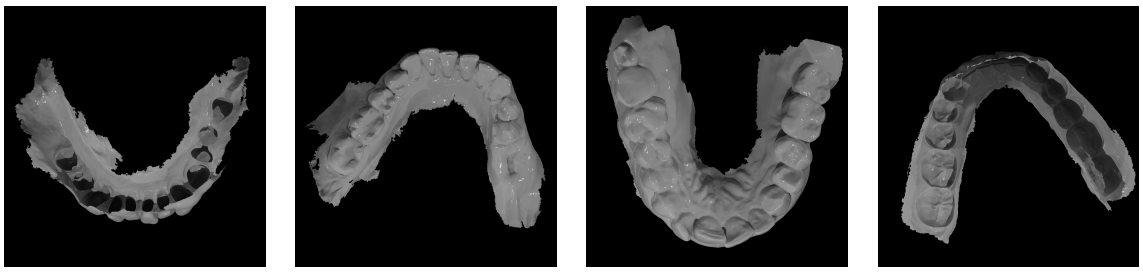
Figure 3.14: **Preliminary test-set renderings under different 3D rotations for the 16+1 jaw-specific model.** The same lower-jaw scan is rendered under multiple rotation configurations. In-plane rotations account for different yaw orientations, while additional rotations around the x or y axis produce tilted views. Although certain tilted views expose additional tooth surfaces, they often distort the global arch structure or reduce overall visibility.

The inspection of these renderings revealed that while some rotated views can locally expose additional tooth surfaces, aggressive out-of-plane rotations frequently lead to reduced global visibility of the dental arch or introduce unstable geometric projections that are less suitable for consistent 2D-based inference. Moreover, incorporating multiple candidate renderings per scan would require an explicit view-selection or multi-view fusion mechanism, which is beyond the scope of the final evaluation protocol adopted in this study.

Based on these observations, we ultimately adopt a simplified test-set rendering strategy for the final evaluation, in which each scan is rendered using a fixed top-view configuration (rot0). This design prioritizes stability and reproducibility while avoiding the complexity of explicit multi-view fusion. All test renderings share the same camera parameters, lighting configuration, and material settings. This design ensures that evaluation differences arise solely from the underlying scan geometry and domain shift, rather than from test-time view selection or multi-angle fusion. Representative examples of the final test-set renderings are shown in Figure 3.15.

However, it is important to emphasize that a fixed top-view rendering does not guarantee semantic orientation consistency across all test samples. Due to the unconstrained nature of real-world intraoral scans, some meshes may still exhibit upside-down configurations, left–right mirroring, or imperfect PCA-based alignment. As a result, even with identical camera parameters, the resulting 2D projections may deviate from an ideal anatomical top-down orientation in certain cases.

This limitation reflects a structural constraint of projection-based 2D pipelines: once a 3D scan is collapsed into a single 2D view, residual orientation errors cannot be corrected without explicit view-selection or multi-view aggregation mechanisms. Consequently, the remaining orientation sensitivity observed in the 2D test results should be interpreted as an inherent limitation of the representation rather than an artifact of rendering parameters.



(a) Lower jaw (b) Lower jaw (c) Upper jaw (d) Upper jaw

Figure 3.15: **Representative test-set renderings using the fixed top view (rot0) for the 16+1 jaw-specific model**

Material and Lighting Design

The rendering configuration employs Physically-Based Rendering (PBR) principles to achieve high-fidelity image quality that enhances diagnostic feature visibility:

Material Shader: We use a unified Principled BSDF (Bidirectional Scattering Distribution Function) shader to ensure a consistent and geometry-focused appearance across all rendered samples.

- **Base Color:** Uniform white, providing a neutral appearance that avoids introducing color bias and emphasizes surface morphology.
- **Specular:** Moderate specularity (0.5) to reflect the semi-glossy characteristics of dental enamel
- **Roughness:** Low roughness (0.1) to preserve sharp surface details while avoiding mirror-like reflections.

To further enhance fine-grained geometric features such as cusps, grooves, and tooth boundaries, an ambient occlusion signal is multiplicatively blended with the base color to enhance local contrast and depth cues. This improves local contrast and depth perception without introducing view-dependent artifacts.

For gap-filling patches introduced during augmentation (e.g., hole filling), the mesh is rendered using the same unified tooth material, ensuring appearance consistency between patched regions and surrounding surfaces.

Lighting Setup: A studio-style multi-light setup is employed to ensure clear, stable, and consistent visualization of tooth anatomy across all rendered samples.

Multiple spot and area lights are distributed around the dental arch and oriented toward a common target using fixed tracking constraints. This configuration provides uniform illumination from multiple directions, reducing harsh shadows and minimizing self-occlusion effects caused by complex tooth geometry.

Rather than relying on a single dominant light source, we employ a symmetric multi-light setup to reduce illumination bias caused by local surface orientation.

This configuration produces a clear visualization of:

- Occlusal surface topography (cusps, grooves, fossae)
- Interproximal spaces (gaps between adjacent teeth)
- Tooth boundaries and contours

Output Specifications

Final renders are generated with the following specifications:

- **Resolution:** 2048×2048 pixels (high resolution to preserve fine geometric details)
- **Color space:** RGBA
- **File format:** PNG (lossless compression)
- **Background:** Uniform black (RGB $= (0, 0, 0)$) for clean segmentation of dental anatomy

In the initial five-view inspection stage, each 3D scan produces five orthographic renderings. This setting is used exclusively for qualitative analysis and viewpoint selection.

For model training and augmentation, only the canonical superior (top) view is retained.

For the test set, different rendering strategies are applied depending on the model formulation: for the 32-neuron model, four in-plane rotations of the top view ($0°, 90°, 180°, 270°$) are generated, while for the jaw-specific 16+1 model, we adopt the final constrained test-time rendering strategy described above, which prioritizes stable global visibility of the dental arch under unconstrained orientations.

The entire rendering process is automated with fixed camera, material, and lighting parameters to ensure consistent outputs across experiments.

3.3.3 Network Architecture

Our 2D classification model is based on the ResNet-18 architecture He et al. (2015), an 18-layer deep residual convolutional network that has demonstrated strong performance on various image classification benchmarks.

ResNet-18 Backbone

The ResNet-18 architecture consists of:

- **Initial convolution:** 7×7 convolution with 64 filters, stride 2, followed by a 3×3 max pooling layer.
- **Residual blocks:** Four subsequent stages of residual blocks (BasicBlocks) with progressively increasing feature dimensions:
  - Stage 1: 2 blocks, 64 filters
  - Stage 2: 2 blocks, 128 filters
  - Stage 3: 2 blocks, 256 filters
  - Stage 4: 2 blocks, 512 filters

- **Global average pooling:** Reduces spatial dimensions to a single 512-dimensional feature vector.

We initialize the backbone with weights pre-trained on ImageNet Deng et al. (2009), which provides strong low-level visual features (edges, textures, shapes). During training, all layers are fine-tuned (not frozen) to adapt features specifically for dental geometry.

Classification Head

The original ImageNet classification head (1000-class softmax) is replaced with a custom multi-label classification head:

1. **Backbone feature extraction:** The global average pooling layer of ResNet-18 produces a 512-dimensional feature vector $f \in \mathbb{R}^{512}$
2. **Fully connected layer:** Linear transformation $f \in \mathbb{R}^{512} \rightarrow z \in \mathbb{R}^{d}$
3. **No activation function:** Output consists of d raw logits

where $d = 32$ for the full-dentition model and $d = 17$ for the jaw-specific $(16 + 1)$ model.

This architecture is formulated as a multi-label classification problem, where each output dimension is predicted independently, reflecting the fact that tooth presence or absence at different positions within a jaw is not mutually exclusive.

3.4 Training and Optimization

3.4.1 Loss Function Design

Baseline-BCE Loss

We formulate automated dental charting as a multi-label binary classification task over 16 tooth positions, where $y_i = 0$ denotes a present tooth and $y_i = 1$ denotes a missing tooth. To establish a performance benchmark, we employ the standard Binary Cross-Entropy (BCE) loss as our baseline. This allows us to evaluate the improvements introduced by our proposed Dynamit loss in handling class imbalance.

Dynamit Loss

To address the severe class imbalance inherent in tooth absence detection (where missing teeth constitute only 5–10% of samples), we employ a dynamic loss function inspired by class-balanced cross-entropy Xie and Tu (2015). This strategy dynamically reweights samples based on the instantaneous positive-negative ratio within each mini-batch, ensuring that the gradient descent is not dominated by the prevalent negative samples (present teeth).

For a given mini-batch, let N_{pos} and N_{neg} denote the total count of positive (missing) and negative (present) samples, respectively. To balance the contribution of each class, we compute adaptive weights w_{pos} and w_{neg} inversely proportional to their batch frequency:

$$w_{\text{pos}} = \min\left(1, \frac{N_{\text{neg}}}{N_{\text{pos}} + \epsilon}\right), \tag{3.2}$$

$$w_{\text{neg}} = \min\left(1, \frac{N_{\text{pos}}}{N_{\text{neg}} + \epsilon}\right), \tag{3.3}$$

where $\epsilon = 10^{-8}$ ensures numerical stability. These weights are assigned point-wise to the binary cross-entropy (BCE) loss. The final objective function is defined as:

$$\mathcal{L}_{\text{Dynamit}} = \frac{1}{B \times d} \sum_{n=1}^{B} \sum_{i=1}^{d} w_{n,i} \cdot \text{BCE}(z_{n,i}, y_{n,i}) \tag{3.4}$$

where $d = 32$ for the full-dentition model and $d = 17$ for the jaw-specific $(16 + 1)$ model. $w_{n,i} \in \{w_{\text{pos}}, w_{\text{neg}}\}$ corresponds to the class of the i-th tooth in sample n, and $z_{n,i}$ represents the raw model logits.

For the jaw-specific $(16 + 1)$ model, dynamic reweighting is applied only to the tooth outputs, while the jaw indicator is trained using standard binary cross-entropy with a fixed weight.

We employ dynamic mini-batch adaptation to balance class contributions. Unlike static methods, this approach automatically reweights samples based on instantaneous batch statistics. This prevents gradient domination by the majority class and prioritizes missing tooth detection, ensuring numerical stability without the need for manual tuning.

Comparison with Alternative Approaches :
Table 3.2 summarizes the characteristics of Dynamit Loss compared to standard alternatives:

Table 3.2: Comparison of loss functions for imbalanced classification

Loss Function	Adaptive Weights	Hyperparameters	Batch-Level
Standard BCE	No	None	–
Weighted BCE	No	Manual class weights	No
Focal Loss Lin et al. (2018)	Yes (by confidence)	γ, α	No
Dynamit Loss	Yes (by batch ratio)	None	Yes

Compared to Focal Loss, which downweights easy-to-classify samples based on prediction confidence, Dynamit Loss directly addresses class imbalance at the sample level. This makes it particularly effective for medical imaging tasks where minority classes (e.g., pathological findings) must be reliably detected despite extreme rarity.

Chapter 4

Experiments and Results

This chapter presents the experimental evaluation of both the 2D render-based and 3D point cloud-based pathways for automated tooth presence classification. We first define the evaluation framework and metrics (§4.1), then present results for each pathway separately (§4.3 and §4.4), followed by systematic ablation studies, and finally a comprehensive comparison between the two approaches.

4.1 Evaluation Framework

4.1.1 Experimental Setup

All models are implemented in PyTorch and trained on GPUs. The 2D pipeline relies on Blender to render each intraoral scan into standardized images. Models are trained for up to 35 or 50 epochs, with early stopping based on validation performance.

4.1.2 Dataset Overview

The dataset is structured based on the FDI World Dental Federation notation, covering 32 distinct tooth positions separated into upper and lower jaws.

Data Distribution

The data cleaning and splitting process resulted in a final dataset structure summarized in Table 4.1. The training set consists of 1800 samples, perfectly balanced between the upper and lower jaws (900 samples each). The test set, after deduplication and pre-processing, comprises 167 unique samples, with 87 images from the upper jaw and 80 from the lower jaw.

Table 4.1: Summary of the dataset distribution after pre-processing.

Subset	Upper Jaw	Lower Jaw	Total Samples
Training Set	900	900	1800
Test Set	87	80	167

Class Imbalance and Tooth Absence

A critical challenge in this dataset is the significant class imbalance caused by naturally missing teeth. Analysis of the test set reveals that third molars (wisdom teeth) exhibit extremely high absence rates compared to other positions.

As detailed in Table 4.2, the absence rates for wisdom teeth (Tooth IDs 18, 28, 38, 48) all exceed 70.0%, with Tooth 28 reaching an absence rate of 85.1%. Furthermore, second molars (e.g., Tooth 17 and 27) also show absence rates above 30.0%, indicating a long-tail distribution that may impact the model's recall performance on these specific classes. All absence statistics are computed with jaw-specific counting to avoid artificial inflation from the opposite jaw.

Table 4.2: Teeth with high absence rates ($\geq 20\%$) in the test set.

Jaw	Tooth ID	Absent Count	Absence Rate (%)
Upper	18	72	82.8
	28	74	85.1
	17	30	34.5
	27	31	35.6
	26	18	20.7
Lower	38	65	81.2
	48	59	73.8
	37	23	28.7
	47	16	20.0
	36	19	23.8
	46	17	21.2
	45	19	23.8

4.1.3 Dataset Splits

The 3DTeethSeg22 dataset Ben-Hamadou et al. (2023b) is split at the case level to avoid duplicate leakage between training and validation samples.

- **Training set:** 1,440 scans (720 patients, 80%)
- **Validation set:** 360 scans (180 patients, 20%)
- **External test set:** 167 unique cases (87 upper jaw and 80 lower jaw) after deduplication and preprocessing

The external test set provides an unbiased assessment of generalization capability.

4.1.4 Evaluation Metrics

To rigorously assess the performance of our models, particularly under conditions of class imbalance, we employ a set of evaluation metrics. We define the standard terms as follows: True Positives (TP) represent correctly identified missing teeth, False Positives (FP) are present teeth incorrectly classified as missing, True Negatives (TN) are correctly identified present teeth, and False Negatives (FN) are missing teeth missed by the model.

- **Precision:** Measures the accuracy of positive predictions (missing teeth).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

- **Recall (Sensitivity):** Measures the proportion of actual missing teeth correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

- **F1-Score:** The harmonic mean of Precision and Recall, providing a single metric that balances both concerns.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

- **Accuracy:** The ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

- **Balanced Accuracy:** Given the extreme class imbalance in our test set (where missing teeth are rare), standard accuracy can be misleading. We therefore include Balanced Accuracy, defined as the arithmetic mean of recall for each class. This metric ensures that the minority class (missing teeth) contributes equally to the final score:

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \tag{4.5}$$

  where K represents the number of classes (typically $K = 2$ for our binary task, effectively averaging Sensitivity and Specificity).

- **Per-tooth metrics:** Precision, Recall, and F1-score are computed independently for each tooth position, while jaw classification performance is evaluated separately for jaw-specific models. These metrics reveal position-specific performance and identify problematic teeth.

4.2 Data Augmentation Results

Our geometry-based augmentation strategy substantially altered the label distribution of the training data, increasing the proportion and variety of missing-tooth patterns. In the original training data, most tooth positions are predominantly present, while a small number of posterior teeth are missing very frequently, which hindered the model's ability to learn negative features.

To address this and improve model robustness, we generated two distinct augmented subsets, bringing the total dataset size to 6,330 samples. The composition of the augmented data is detailed below and summarized in Table 4.3.

4.2.1 Augmentation Strategies and Distribution

The augmentation pipeline employed two different sampling strategies to serve distinct training and validation objectives:

1. Random Augmentation (Training Support) The largest subset, consisting of **3,600 samples** (1,800 upper jaw, 1,800 lower jaw), was generated using a stochastic removal strategy. In this process, Teeth are removed via weighted sampling to increase pattern diversity, while down-weighting teeth that are already frequently absent, so the augmented set does not further skew the distribution. This strategy forces the network to learn tooth features in isolation rather than relying on the presence of neighboring teeth, effectively mitigating the severe class imbalance inherent in the raw data.

2. Distribution-based Augmentation (Robustness Testing) A second subset of **930 samples** (485 upper jaw, 445 lower jaw) was generated as a robustness-focused set. It is guided by test-set statistics and clinically plausible constraints, but remains grounded in the training scans, thereby retaining naturally high absence for certain posterior teeth. This ensures that the model is evaluated and validated on data that closely resembles the statistical properties of the final inference targets.

Table 4.3: Summary of the total dataset composition after augmentation.

Dataset Component	Strategy	Split (Upper/Lower)	Total Samples
Original Train Set	Raw Data	900 / 900	1800
Augmented Set A	Random Removal	1800 / 1800	3600
Augmented Set B	Test-Distribution	485 / 445	930
Grand Total	**-**	**3185 / 3145**	**6330**

Quantitative analysis confirms that the distribution-based subset stabilizes the average missing rate at 39.0% ($\sigma = 22.2\%$). This strategy effectively mitigates imbalance while preserving clinical realism, such as retaining naturally high absence rates ($> 90\%$) for wisdom teeth (e.g., Tooth 28).

A comparative analysis of the two augmentation strategies reveals distinct distributional characteristics. The random augmentation strategy (3,600 samples) generally yields a slightly more conservative missingness profile compared to the distribution-based approach. Specifically, 78.1% of tooth positions (25 out of 32) exhibit fewer missing samples in the random subset, with an average deviation of -2.7% in missing rates. This suggests that while the random strategy maximizes pattern diversity through volume, the distribution-based strategy more strongly reflects clinically realistic high-absence rates observed in posterior teeth.

4.3 2D Render-Based Pathway Results

4.3.1 2D Results (16+1 Architecture)

The $16 + 1$ architecture distinguishes between upper and lower jaws. The results are filtered for Support > 0. Support denotes the number of missing-tooth positives (label = 1) at each tooth position. As a result, metrics (Precision, Recall, F1, and Accuracy) are not available for tooth positions with no positive instances. Jaw accuracy is computed solely on the jaw neuron, independent of per-tooth outputs.

Training Phase Results

Comparison of loss functions and augmentation strategies. **Hybrid Aug** refers to Original + Test-Pattern + Random Synthetic data.

Tables 4.4 and 4.5 summarize performance on the validation split for the 16+1 architecture (used for early stopping). Table 4.4 reports macro-averaged tooth-level metrics over positions with Support > 0, computed on the 16 tooth outputs only (excluding the jaw neuron), while Table 4.5 reports accuracy on the additional jaw neuron. Due to extreme class imbalance, accuracy can remain high even when missing-tooth recall is limited. Therefore, recall and F1 are more indicative for the missing-tooth detection objective. Compared to the BCE baseline, Dynamit improves tooth-level recall by dynamically reweighting positives within each mini-batch. Incorporating hybrid augmentation (original scans + balanced synthetic removals + random synthetic removals) substantially increases positive support and yields more stable per-tooth estimates on the validation split. Jaw accuracy remains consistently high across configurations, as the jaw neuron is trained with standard BCE and evaluated independently.

Table 4.4: **Overall Tooth Detection Performance (Support > 0)**.

Model Configuration	Precision	Recall	F1 Score	Accuracy
Baseline (BCE Loss)	0.2906	0.2972	0.2932	0.9692
Dynamit (Original Data)	0.3772	0.4556	0.4054	0.9621
Dynamit + Hybrid Aug (Orig+Target+Rand)	**0.9428**	**0.9554**	**0.9487**	**0.9754**

Table 4.5: **Jaw Classification Accuracy**. Evaluation of the model's ability to distinguish between upper and lower jaws (the +1 neuron) across different training configurations.

Model Configuration	Jaw Accuracy
Baseline (BCE Loss)	0.9889
Dynamit (Original Data)	0.9944
Dynamit + Hybrid Aug (Orig+Target+Rand)	**0.9976**

Table 4.6, Table 4.7, and Table 4.8 report per-tooth training performance as a diagnostic breakdown. Many tooth positions have very few or zero missing-tooth positives in the original training set. Therefore per-tooth PRF1 can be undefined (Support = 0) or unstable for rare positions. Augmentation increases positive support and yields more reliable per-tooth estimates.

Table 4.6: **2D 16+1 Baseline (BCE) Training Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support (number of positive samples).

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.9766	0.9882	0.9824	0.9667	169	167	4	2	7	48	0.9595	0.9822	0.9708	0.9444	169	166	7	3	4
17	0.9057	0.9412	0.9231	0.9556	51	48	5	3	124	47	0.8367	0.8913	0.8632	0.9278	46	41	8	5	126
16	0.0000	0.0000	0.0000	0.9889	2	0	0	2	178	46	0.0000	0.0000	0.0000	0.9889	2	0	0	2	178
15	0.0000	0.0000	0.0000	0.9722	5	0	0	5	175	45	0.0000	0.0000	0.0000	0.9667	6	0	0	6	174
14	0.0000	0.0000	0.0000	0.9889	2	0	0	2	178	44	0.0000	0.0000	0.0000	0.9833	3	0	0	3	177
13	0.5385	0.5000	0.5185	0.9278	14	7	6	7	160	43	0.0000	0.0000	0.0000	0.9889	1	0	1	1	178
12	0.0000	0.0000	0.0000	0.9667	6	0	0	6	174	42	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
11	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179	41	–	–	–	1.0000	0	0	0	0	180
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179	31	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
22	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179	32	–	–	–	1.0000	0	0	0	0	180
23	0.7000	0.5385	0.6087	0.9500	13	7	3	6	164	33	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
24	0.0000	0.0000	0.0000	0.9778	4	0	0	4	176	34	0.0000	0.0000	0.0000	0.9722	5	0	0	5	175
25	0.0000	0.0000	0.0000	0.9722	5	0	0	5	175	35	0.0000	0.0000	0.0000	0.9778	4	0	0	4	176
26	–	–	–	1.0000	0	0	0	0	180	36	0.0000	0.0000	0.0000	0.9833	3	0	0	3	177
27	0.8103	0.9400	0.8704	0.9222	50	47	11	3	119	37	0.7708	0.8605	0.8132	0.9056	43	37	11	6	126
28	0.9767	0.9882	0.9825	0.9667	170	168	4	2	6	38	0.9535	0.9880	0.9704	0.9444	166	164	8	2	6

Table 4.7: **2D 16+1 Dynamit Loss Training Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support (number of positive samples).

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.9389	1.0000	0.9685	0.9389	169	169	11	0	0	48	0.9494	1.0000	0.9741	0.9500	169	169	9	0	2
17	0.8545	0.9216	0.8868	0.9333	51	47	8	4	121	47	0.8000	0.8696	0.8333	0.9111	46	40	10	6	124
16	0.5000	0.5000	0.5000	0.9889	2	1	1	1	177	46	0.1667	0.5000	0.2500	0.9667	2	1	5	1	173
15	0.2857	0.4000	0.3333	0.9556	5	2	5	3	170	45	0.1429	0.1667	0.1538	0.9389	6	1	6	5	168
14	0.2000	0.5000	0.2857	0.9722	2	1	4	1	174	44	0.0000	0.0000	0.0000	0.9833	3	0	0	3	177
13	0.5294	0.6429	0.5806	0.9278	14	9	8	5	158	43	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
12	0.0000	0.0000	0.0000	0.9667	6	0	0	6	174	42	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
11	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179	41	–	–	–	1.0000	0	0	0	0	180
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179	31	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
22	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179	32	–	–	–	1.0000	0	0	0	0	180
23	0.6250	0.7692	0.6897	0.9500	13	10	6	3	161	33	0.0000	0.0000	0.0000	0.9944	1	0	0	1	179
24	0.0000	0.0000	0.0000	0.9556	4	0	4	4	172	34	0.3333	0.2000	0.2500	0.9667	5	1	2	4	173
25	0.2500	0.4000	0.3077	0.9500	5	2	6	3	169	35	0.2222	0.5000	0.3077	0.9500	4	2	7	2	169
26	–	–	–	0.9833	0	0	3	0	177	36	0.6000	1.0000	0.7500	0.9889	3	3	2	0	175
27	0.8727	0.9600	0.9143	0.9500	50	48	7	2	123	37	0.7917	0.8837	0.8352	0.9167	43	38	10	5	127
28	0.9444	1.0000	0.9714	0.9444	170	170	10	0	0	38	0.9326	1.0000	0.9651	0.9333	166	166	12	0	2

Table 4.8: **2D 16+1 Dynamit + Hybrid Augmentation Training Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). Hybrid augmentation includes original, targeted synthetic, and random synthetic data. *Supp* indicates Support.

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.9901	0.9983	0.9942	0.9889	602	601	6	1	22	48	0.9851	0.9900	0.9876	0.9763	601	595	9	6	22
17	0.9167	0.9735	0.9442	0.9587	226	220	20	6	384	47	0.9183	0.9409	0.9294	0.9541	203	191	17	12	412
16	0.9130	0.9375	0.9251	0.9730	112	105	10	7	508	46	0.9187	0.9187	0.9187	0.9684	123	113	10	10	499
15	0.8963	0.9030	0.8996	0.9571	134	121	14	13	482	45	0.9091	0.9302	0.9195	0.9668	129	120	12	9	491
14	0.9635	0.9429	0.9531	0.9794	140	132	5	8	485	44	0.9474	0.9558	0.9515	0.9826	113	108	6	5	513
13	0.9295	0.9355	0.9325	0.9667	155	145	11	10	464	43	0.9545	0.9545	0.9545	0.9810	132	126	6	6	494
12	0.9621	0.9407	0.9513	0.9794	135	127	5	8	490	42	0.9364	0.9450	0.9406	0.9794	109	103	7	6	516
11	0.9624	0.9922	0.9771	0.9905	129	128	5	1	496	41	0.9462	0.9609	0.9535	0.9810	128	123	7	5	497
Left Quadrant (Q2 vs Q3)																			
21	0.9573	0.9912	0.9739	0.9905	113	112	5	1	512	31	0.8913	0.9919	0.9389	0.9747	124	123	15	1	493
22	0.9573	0.9655	0.9614	0.9857	116	112	5	4	509	32	0.9412	0.9739	0.9573	0.9842	115	112	7	3	510
23	0.9706	0.9565	0.9635	0.9841	138	132	4	6	488	33	0.9583	0.9200	0.9388	0.9763	125	115	5	10	502
24	0.9720	0.9456	0.9586	0.9810	147	139	4	8	479	34	0.9677	0.9449	0.9562	0.9826	127	120	4	7	501
25	0.8375	0.9504	0.8904	0.9476	141	134	26	7	463	35	0.9000	0.9333	0.9164	0.9636	135	126	14	9	483
26	0.9520	0.9520	0.9520	0.9810	125	119	6	6	499	36	0.9746	0.9274	0.9504	0.9810	124	115	3	9	505
27	0.9174	0.9724	0.9441	0.9603	217	211	19	6	394	37	0.9372	0.9372	0.9372	0.9620	191	179	12	12	429
28	0.9918	0.9967	0.9942	0.9889	605	603	5	2	20	38	0.9934	0.9934	0.9934	0.9873	604	600	4	4	24

Testing Phase Results

During evaluation, only cases with successfully rendered 2D images and valid CSV labels were retained. After alignment and filtering, the final test set consisted of 167 cases (87 upper and 80 lower jaws), which were consistently used across all experiments.

Tables 4.9 and Table 4.10 summarize the overall test performance across different training strategies. The extreme class imbalance in the test set means that standard accuracy alone can be misleading, as correctly predicting the dominant "present" class yields high accuracy even when missing teeth are frequently misclassified. Therefore, we additionally report macro-averaged precision, recall, F1 score, and balanced accuracy, which better reflect performance on both missing and present classes. Precision/Recall/F1 are macro-averaged across tooth positions, whereas the confusion matrix reports aggregated counts over all teeth.

Table 4.9: **Confusion Matrix Comparison (Teeth Missing/Present).** *TN*: True Negatives (Correctly predicted present teeth), *FP*: False Positives (False alarms), *FN*: False Negatives (Missed missing teeth), *TP*: True Positives (Correctly detected missing teeth).

Model Configuration	**TN**	**FP**	**FN**	**TP**
Baseline (BCE Loss)	2021	89	247	315
Dynamit (Original Data)	1911	199	193	369
Dynamit + Hybrid Aug (Orig+Target+Rand)	1835	275	206	356

Compared to BCE, Dynamit increases true positives (315→369) and reduces false negatives (247→193), indicating improved sensitivity to missing teeth, at the cost of more false positives (89→199). Hybrid augmentation further increases precision (Table 4.10) but continues to trade off specificity via additional false positives (275).

Table 4.10: **Overall Performance Metrics.**

Model Configuration	Precision	Recall	F1 Score	Accuracy	Bal. Acc.
Baseline (BCE Loss)	0.1869	0.1873	0.1851	**0.8748**	0.5356
Dynamit (Original Data)	0.2522	0.2905	0.2614	0.8536	0.5547
Dynamit + Hybrid Aug	**0.3092**	**0.3101**	**0.2873**	0.8220	**0.5713**

Table 4.11: **Jaw Classification Accuracy.** Evaluation of the auxiliary neuron (+1) determining whether the input scan is Upper or Lower jaw.

Model Configuration	Jaw Accuracy
Baseline (BCE Loss)	0.7545
Dynamit (Original Data)	**0.9042**
Dynamit + Hybrid Aug (Orig+Target+Rand)	0.8802

Although BCE achieves the highest raw accuracy, its balanced accuracy and macro-F1 are lower, consistent with a bias toward predicting the dominant present class. Hybrid augmentation improves macro-level precision/recall and balanced accuracy, indicating better class-balanced behavior despite a drop in overall accuracy.

Table 4.11 reports the jaw classification accuracy of the auxiliary output neuron. Dynamit substantially improves jaw prediction compared to the BCE baseline. Adding hybrid augmentation does not further improve jaw accuracy in this evaluation, suggesting that gains are not monotonic across training strategies. Note that jaw accuracy is evaluated on the dedicated jaw neuron trained with standard BCE (fixed weight), independently of the dynamically reweighted tooth losses.

Tables (Table 4.12, Table 4.13, and Table 4.14 present a per-tooth performance breakdown for both upper and lower jaws. These results provide a fine-grained view of model behavior across individual teeth. However, per-tooth results exhibit high variance for teeth with small support, where a small number of errors can substantially change precision/recall. Therefore, these tables are primarily used as diagnostic evidence to identify systematically difficult positions.

Table 4.12: **16+1 Baseline (BCE) Test Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support (number of positive samples).

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.8750	0.9722	0.9211	0.8621	72	70	10	2	5	48	0.7632	0.9831	0.8593	0.7625	59	58	18	1	3
17	0.7619	0.5333	0.6275	0.7816	30	16	5	14	52	47	0.5000	0.5000	0.5000	0.8000	16	8	8	8	56
16	0.0000	0.0000	0.0000	0.8161	16	0	0	16	71	46	0.0000	0.0000	0.0000	0.7875	17	0	0	17	63
15	0.0000	0.0000	0.0000	0.8736	11	0	0	11	76	45	0.0000	0.0000	0.0000	0.7625	19	0	0	19	61
14	0.0000	0.0000	0.0000	0.8621	12	0	0	12	75	44	0.0000	0.0000	0.0000	0.9375	5	0	0	5	75
13	0.0000	0.0000	0.0000	0.9080	1	0	7	1	79	43	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
12	0.0000	0.0000	0.0000	0.9310	6	0	0	6	81	42	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
11	0.0000	0.0000	0.0000	0.9195	7	0	0	7	80	41	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9425	5	0	0	5	82	31	0.0000	0.0000	0.0000	0.9750	2	0	0	2	78
22	0.0000	0.0000	0.0000	0.9540	4	0	0	4	83	32	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
23	0.0000	0.0000	0.0000	0.9195	3	0	4	3	80	33	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
24	0.0000	0.0000	0.0000	0.8621	12	0	0	12	75	34	0.0000	0.0000	0.0000	0.9500	4	0	0	4	76
25	0.0000	0.0000	0.0000	0.8506	13	0	0	13	74	35	0.0000	0.0000	0.0000	0.8875	9	0	0	9	71
26	0.0000	0.0000	0.0000	0.7931	18	0	0	18	69	36	0.0000	0.0000	0.0000	0.7625	19	0	0	19	61
27	0.5385	0.4516	0.4912	0.6667	31	14	12	17	44	37	0.8235	0.6087	0.7000	0.8500	23	14	3	9	54
28	0.8659	0.9595	0.9103	0.8391	74	71	11	3	2	38	0.8533	0.9846	0.9143	0.8500	65	64	11	1	4

Table 4.13: **2D 16+1 Dynamit Test Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). Results are reordered from the source to match standard FDI quadrant notation. *Supp* indicates Support.

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.8235	0.9722	0.8917	0.8046	72	70	15	2	0	48	0.7733	0.9831	0.8657	0.7750	59	58	17	1	4
17	0.7895	0.5000	0.6122	0.7816	30	15	4	15	53	47	0.3750	0.7500	0.5000	0.7000	16	12	20	4	44
16	0.4444	0.2500	0.3200	0.8046	16	4	5	12	66	46	0.5000	0.7647	0.6047	0.7875	17	13	13	4	50
15	0.1429	0.0909	0.1111	0.8161	11	1	6	10	70	45	0.1667	0.0526	0.0800	0.7125	19	1	5	18	56
14	0.0000	0.0000	0.0000	0.8391	12	0	2	12	73	44	0.0000	0.0000	0.0000	0.9375	5	0	0	5	75
13	0.0000	0.0000	0.0000	0.9425	1	0	4	1	82	43	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
12	0.0000	0.0000	0.0000	0.9310	6	0	0	6	81	42	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
11	0.0000	0.0000	0.0000	0.9195	7	0	0	7	80	41	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9425	5	0	0	5	82	31	0.0000	0.0000	0.0000	0.9750	2	0	0	2	78
22	0.0000	0.0000	0.0000	0.9540	4	0	0	4	83	32	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
23	0.0000	0.0000	0.0000	0.9080	3	0	5	3	79	33	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
24	0.0000	0.0000	0.0000	0.8391	12	0	2	12	73	34	0.0000	0.0000	0.0000	0.9500	4	0	0	4	76
25	0.2000	0.1538	0.1739	0.7816	13	2	8	11	66	35	0.2381	0.5556	0.3333	0.7500	9	5	16	4	55
26	0.3333	0.2778	0.3030	0.7356	18	5	10	13	59	36	0.4783	0.5789	0.5238	0.7500	19	11	12	8	49
27	0.5938	0.6129	0.6032	0.7126	31	19	13	12	43	37	0.5278	0.8261	0.6441	0.7375	23	19	17	4	40
28	0.8452	0.9595	0.8987	0.8161	74	71	13	3	0	38	0.8400	0.9692	0.9000	0.8250	65	63	12	2	3

Table 4.14: **16+1 Dynamit + Hybrid Augmentation Test Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). Results are reordered to match standard FDI quadrant notation. *Supp* indicates Support.

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.8667	0.9028	0.8844	0.8046	72	65	10	7	5	48	0.8308	0.9153	0.8710	0.8000	59	54	11	5	10
17	0.8667	0.4333	0.5778	0.7816	30	13	2	17	55	47	0.3333	0.2500	0.2857	0.7500	16	4	8	12	56
16	0.2692	0.4375	0.3333	0.6782	16	7	19	9	52	46	0.4000	0.3529	0.3750	0.7500	17	6	9	11	54
15	0.1795	0.6364	0.2800	0.5862	11	7	32	4	44	45	0.6250	0.2632	0.3704	0.7875	19	5	3	14	58
14	0.4000	0.5000	0.4444	0.8276	12	6	9	6	66	44	0.0000	0.0000	0.0000	0.9375	5	0	0	5	75
13	0.0000	0.0000	0.0000	0.7701	1	0	19	1	67	43	0.0000	0.0000	0.0000	0.9750	1	0	1	1	78
12	0.5000	0.1667	0.2500	0.9310	6	1	1	5	80	42	0.0000	0.0000	0.0000	0.9500	3	0	1	3	76
11	0.1429	0.1429	0.1429	0.8621	7	1	6	6	74	41	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9195	5	0	2	5	80	31	0.0000	0.0000	0.0000	0.9625	2	0	1	2	77
22	0.0000	0.0000	0.0000	0.9310	4	0	2	4	81	32	0.0000	0.0000	0.0000	0.9750	1	0	1	1	78
23	0.0000	0.0000	0.0000	0.7931	3	0	15	3	69	33	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
24	0.0769	0.0833	0.0800	0.7356	12	1	12	11	63	34	0.0000	0.0000	0.0000	0.9500	4	0	0	4	76
25	0.1818	0.7692	0.2941	0.4483	13	10	45	3	29	35	0.1176	0.2222	0.1538	0.7250	9	2	15	7	56
26	0.3611	0.7222	0.4815	0.6782	18	13	23	5	46	36	0.4286	0.3158	0.3636	0.7375	19	6	8	13	53
27	0.8000	0.5161	0.6275	0.7816	31	16	4	15	52	37	0.6667	0.4348	0.5263	0.7750	23	10	5	13	52
28	0.8947	0.9189	0.9067	0.8391	74	68	8	6	5	38	0.9531	0.9385	0.9457	0.9125	65	61	3	4	12

4.4 3D Point Cloud-Based Pathway Results

4.4.1 3D Results (16+1 Architecture)

Training Phase Results

Table 4.15 reports the macro-averaged training performance of the 3D 16+1 architecture under three different optimization strategies. The Baseline (BCE) configuration achieves relatively low Precision (0.4395), Recall (0.3815), and F1-score (0.3906), despite a high overall Accuracy of 0.9629. Dynamit loss increases Recall to 0.4537 and F1-score to 0.4261, while Precision remains comparable to the baseline.

A substantial improvement is observed when Hybrid Augmentation dataset is used. The Dynamit + Hybrid Aug configuration achieves higher Precision (0.8915), Recall (0.9107), and F1-score (0.9006), indicating strong convergence across tooth-level predictions during training. Accuracy remains high (0.9520), comparable to the non-augmented settings.

Table 4.15: **Comparison of 3D 16+1 Training Metrics.** Evaluated on the training set. *Baseline* and *Dynamit* use original data only, while *Hybrid Aug* incorporates synthetic data, resulting in significantly higher convergence metrics.

Model Configuration	Precision	Recall	F1 Score	Accuracy
Baseline (BCE Loss)	0.4395	0.3815	0.3906	**0.9629**
Dynamit	0.4381	0.4537	0.4261	0.9526
Dynamit + Hybrid Aug	**0.8915**	**0.9107**	**0.9006**	0.9520

Table 4.16 summarizes the auxiliary jaw classification accuracy on the training set. All config-

urations achieve near-perfect performance. The Baseline model attains an accuracy of 0.9944, which further increases to 0.9972 under Dynamit loss. The highest jaw classification accuracy is achieved by the Dynamit + Hybrid Aug model at 0.9992, indicating highly reliable separation between upper and lower jaws during training.

Table 4.16: **3D 16+1 Jaw Classification Accuracy (Training Set).** Comparison of the auxiliary task (Upper vs. Lower jaw classification) across different training strategies.

Model Configuration	Jaw Accuracy
Baseline (BCE Loss)	0.9944
Dynamit	0.9972
Dynamit + Hybrid Aug	**0.9992**

Tables 4.17, 4.18, and 4.19 present a comparison of per-tooth training performance for the 3D 16+1 architecture under three learning strategies: Baseline (BCE), Dynamit loss, and Dynamit with Hybrid Augmentation.

The Baseline model exhibits a strong bias toward high-support posterior teeth, achieving near-perfect performance for molars (e.g., FDI 18/48, 28/38), while most anterior teeth collapse to zero Recall despite high Accuracy driven by true negatives. The Dynamit loss partially alleviates this imbalance, improving Recall for several low-frequency teeth, but performance remains unstable in anterior regions. In contrast, Dynamit combined with Hybrid Augmentation yields high Precision and Recall across all tooth positions, including anterior incisors, with substantially improved F1-scores. These results demonstrate that loss reweighting alone is insufficient, and that distribution-aware data augmentation is essential for stable per-tooth learning under extreme class imbalance.

Table 4.17: **3D 16+1 Baseline (BCE) Training Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support (number of positive samples).

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.9773	0.9718	0.9745	0.9508	177	172	4	5	2	48	0.9588	0.9879	0.9731	0.9492	165	163	7	2	5
17	0.8475	0.9091	0.8772	0.9235	55	50	9	5	119	47	0.8065	0.6757	0.7353	0.8983	37	25	6	12	134
16	–	–	–	1.0000	0	0	0	0	183	46	–	–	–	0.9944	0	0	1	0	176
15	1.0000	0.2000	0.3333	0.9781	5	1	0	4	178	45	0.3333	0.3333	0.3333	0.9774	3	1	2	2	172
14	0.0000	0.0000	0.0000	0.9727	5	0	0	5	178	44	0.0000	0.0000	0.0000	0.9774	4	0	0	4	173
13	0.6667	0.2500	0.3636	0.9235	16	4	2	12	165	43	0.2500	0.5000	0.3333	0.9774	2	1	3	1	172
12	0.0000	0.0000	0.0000	0.9781	4	0	0	4	179	42	–	–	–	1.0000	0	0	0	0	177
11	–	–	–	1.0000	0	0	0	0	183	41	0.0000	0.0000	0.0000	0.9944	1	0	0	1	176
Left Quadrant (Q2 vs Q3)																			
21	–	–	–	1.0000	0	0	0	0	183	31	0.0000	0.0000	0.0000	0.9887	2	0	0	2	175
22	0.0000	0.0000	0.0000	0.9945	1	0	0	1	182	32	–	–	–	1.0000	0	0	0	0	177
23	0.8889	0.7273	0.8000	0.9781	11	8	1	3	171	33	0.6667	0.5000	0.5714	0.9831	4	2	1	2	172
24	0.0000	0.0000	0.0000	0.9945	1	0	0	1	182	34	0.0000	0.0000	0.0000	0.9774	4	0	0	4	173
25	0.0000	0.0000	0.0000	0.9563	8	0	0	8	175	35	0.5000	0.1667	0.2500	0.9661	6	1	1	5	170
26	0.0000	0.0000	0.0000	0.9836	3	0	0	3	180	36	0.0000	0.0000	0.0000	0.9774	4	0	0	4	173
27	0.8000	0.8302	0.8148	0.8907	53	44	11	9	119	37	0.7895	0.9091	0.8451	0.9379	33	30	8	3	136
28	0.9831	0.9775	0.9803	0.9617	178	174	3	4	2	38	0.9586	0.9818	0.9701	0.9435	165	162	7	3	5

Table 4.18: **3D 16+1 Dynamit Training Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support (number of positive samples).

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.9672	1.0000	0.9833	0.9672	177	177	6	0	0	48	0.9322	1.0000	0.9649	0.9322	165	165	12	0	0
17	0.7538	0.8909	0.8167	0.8798	55	49	16	6	112	47	0.6750	0.7297	0.7013	0.8701	37	27	13	10	127
16	–	–	–	1.0000	0	0	0	0	183	46	–	–	–	0.9887	0	0	2	0	175
15	0.5000	0.2000	0.2857	0.9727	5	1	1	4	177	45	0.1000	0.3333	0.1538	0.9379	3	1	9	2	165
14	0.5000	0.2000	0.2857	0.9727	5	1	1	4	177	44	0.0000	0.0000	0.0000	0.9774	4	0	0	4	173
13	0.4500	0.5625	0.5000	0.9016	16	9	11	7	156	43	0.2000	0.5000	0.2857	0.9718	2	1	4	1	171
12	0.0000	0.0000	0.0000	0.9727	4	0	1	4	178	42	–	–	–	1.0000	0	0	0	0	177
11	–	–	–	1.0000	0	0	0	0	183	41	0.0000	0.0000	0.0000	0.9944	1	0	0	1	176
Left Quadrant (Q2 vs Q3)																			
21	–	–	–	1.0000	0	0	0	0	183	31	0.0000	0.0000	0.0000	0.9887	2	0	0	2	175
22	0.0000	0.0000	0.0000	0.9945	1	0	0	1	182	32	–	–	–	1.0000	0	0	0	0	177
23	0.4706	0.7273	0.5714	0.9344	11	8	9	3	163	33	0.5000	0.7500	0.6000	0.9774	4	3	3	1	170
24	0.0000	0.0000	0.0000	0.9945	1	0	0	1	182	34	0.0000	0.0000	0.0000	0.9605	4	0	3	4	170
25	0.0000	0.0000	0.0000	0.9399	8	0	3	8	172	35	0.2857	0.3333	0.3077	0.9492	6	2	5	4	166
26	1.0000	0.3333	0.5000	0.9891	3	1	0	2	180	36	0.6667	0.5000	0.5714	0.9831	4	2	1	2	172
27	0.7833	0.8868	0.8319	0.8962	53	47	13	6	117	37	0.7000	0.8485	0.7671	0.9040	33	28	12	5	132
28	0.9727	1.0000	0.9861	0.9727	178	178	5	0	0	38	0.9322	1.0000	0.9649	0.9322	165	165	12	0	0

Table 4.19: **3D 16+1 Dynamit + Hybrid Augmentation Training Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). Hybrid augmentation includes original, targeted synthetic, and random synthetic data. *Supp* indicates Support.

Upper Jaw (Maxillary)						Lower Jaw (Mandibular)					
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**
Right Quadrant (Q1 vs Q4)											
18	0.9604	0.9967	0.9782	0.9580	608	48	0.9752	0.9949	0.9849	0.9711	592
17	0.8782	0.9084	0.8931	0.9114	262	47	0.8068	0.8978	0.8499	0.9053	186
16	0.8723	0.9318	0.9011	0.9580	132	46	0.9035	0.8729	0.8879	0.9583	118
15	0.8672	0.8346	0.8506	0.9393	133	45	0.8087	0.7949	0.8017	0.9262	117
14	0.8732	0.8857	0.8794	0.9471	140	44	0.8814	0.8667	0.8739	0.9518	120
13	0.8431	0.8897	0.8658	0.9378	145	43	0.9160	0.9237	0.9198	0.9695	118
12	0.9371	0.9241	0.9306	0.9689	145	42	0.9217	0.8908	0.9060	0.9647	119
11	0.9697	0.9771	0.9734	0.9891	131	41	0.8667	0.9123	0.8889	0.9583	114
Left Quadrant (Q2 vs Q3)											
21	0.9565	1.0000	0.9778	0.9922	110	31	0.9043	0.8966	0.9004	0.9631	116
22	0.9134	0.9431	0.9280	0.9720	123	32	0.9231	0.8972	0.9100	0.9695	107
23	0.8491	0.8882	0.8682	0.9362	152	33	0.8881	0.9154	0.9015	0.9583	130
24	0.8931	0.8931	0.8931	0.9565	131	34	0.8621	0.8929	0.8772	0.9551	112
25	0.8538	0.7986	0.8253	0.9269	139	35	0.7686	0.8532	0.8087	0.9294	109
26	0.9492	0.9180	0.9333	0.9751	122	36	0.8837	0.9500	0.9157	0.9663	120
27	0.8365	0.9167	0.8748	0.9020	240	37	0.8204	0.8802	0.8492	0.9037	192
28	0.9701	0.9984	0.9840	0.9689	617	38	0.9735	0.9983	0.9858	0.9727	589

Testing Phase Results

Table 4.20 summarizes the macro-averaged performance of the three 3D 16+1 model variants on the held-out test set.

The baseline model trained with BCE loss achieves an overall accuracy of 0.8534, while showing relatively low recall (0.2047) and F1 score (0.1892), indicating limited sensitivity to missing-tooth cases. Replacing the BCE loss with Dynamit improves recall to 0.2906 and F1 score to 0.2315, with a slight decrease in overall accuracy to 0.8339.

The Dynamit model trained with hybrid augmentation yields the strongest performance in terms of minority-class detection, achieving the highest precision (0.3769), recall (0.5565), and F1 score (0.3743). Although its overall accuracy (0.7994) is lower than that of the baseline, it attains the highest balanced accuracy (0.7707), reflecting improved performance across both present and missing tooth classes.

Table 4.20: **Comparison of 3D 16+1 Test Metrics.** Evaluated on the held-out test set. *Dynamit + Hybrid Augmentation* achieves the highest Recall and F1 score, demonstrating better generalization on minority classes (missing teeth) despite a lower standard accuracy compared to the baseline.

Model Configuration	Precision	Recall	F1 Score	Accuracy	Bal. Acc.
Baseline (BCE Loss)	0.2360	0.2047	0.1892	0.8534	0.7524
Dynamit	0.2360	0.2906	0.2315	0.8339	0.7531
Dynamit + Hybrid Aug	**0.3769**	**0.5565**	**0.3743**	0.7994	**0.7707**

Jaw-level classification accuracy for the auxiliary upper/lower jaw prediction task is reported in Table 4.21.

The baseline BCE model achieves the highest jaw classification accuracy of 0.9341. The Dynamit model trained on the original data shows a comparable performance at 0.9281. In contrast, the Dynamit model with hybrid augmentation exhibits a reduced jaw accuracy of 0.8982, despite its superior tooth-level detection performance.

Table 4.21: **3D 16+1 Jaw Classification Accuracy (Test Set).** Comparison of the auxiliary task (Upper vs. Lower jaw classification) performance on the test set.

Model Configuration	Jaw Accuracy
Baseline (BCE Loss)	**0.9341**
Dynamit (Original Data)	0.9281
Dynamit + Hybrid Aug	0.8982

Table 4.22: **3D 16+1 Baseline (BCE) Test Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support.

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.8276	1.0000	0.9057	0.8276	72	72	15	0	0	48	0.7375	1.0000	0.8489	0.7375	59	59	21	0	0
17	0.4706	0.2667	0.3404	0.6437	30	8	9	22	48	47	0.3810	0.5000	0.4324	0.7375	16	8	13	8	51
16	0.3333	0.0625	0.1053	0.8046	16	1	2	15	69	46	0.3158	0.3529	0.3333	0.7000	17	6	13	11	50
15	0.0000	0.0000	0.0000	0.8391	11	0	3	11	73	45	0.0000	0.0000	0.0000	0.7625	19	0	0	19	61
14	0.0000	0.0000	0.0000	0.8621	12	0	0	12	75	44	0.0000	0.0000	0.0000	0.9375	5	0	0	5	75
13	0.0000	0.0000	0.0000	0.9885	1	0	0	1	86	43	0.0000	0.0000	0.0000	0.9500	1	0	3	1	76
12	0.0000	0.0000	0.0000	0.9310	6	0	0	6	81	42	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
11	0.0000	0.0000	0.0000	0.9195	7	0	0	7	80	41	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9425	5	0	0	5	82	31	0.0000	0.0000	0.0000	0.9750	2	0	0	2	78
22	0.0000	0.0000	0.0000	0.9540	4	0	0	4	83	32	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
23	0.0000	0.0000	0.0000	0.9540	3	0	1	3	83	33	0.0000	0.0000	0.0000	0.9375	1	0	4	1	75
24	0.0000	0.0000	0.0000	0.8621	12	0	0	12	75	34	0.0000	0.0000	0.0000	0.9500	4	0	0	4	76
25	1.0000	0.0769	0.1429	0.8621	13	1	0	12	74	35	1.0000	0.1111	0.2000	0.9000	9	1	0	8	71
26	0.0000	0.0000	0.0000	0.7931	18	0	0	18	69	36	0.0000	0.0000	0.0000	0.7625	19	0	0	19	61
27	0.4074	0.3548	0.3793	0.5862	31	11	16	20	40	37	0.4043	0.8261	0.5429	0.6000	23	19	28	4	29
28	0.8506	1.0000	0.9193	0.8506	74	74	13	0	0	38	0.8228	1.0000	0.9028	0.8250	65	65	14	0	1

Table 4.23: **3D 16+1 Dynamit Test Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support.

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.8276	1.0000	0.9057	0.8276	72	72	15	0	0	48	0.7375	1.0000	0.8489	0.7375	59	59	21	0	0
17	0.4848	0.5333	0.5079	0.6437	30	16	17	14	40	47	0.4286	0.1875	0.2609	0.7875	16	3	4	13	60
16	0.4286	0.1875	0.2609	0.8046	16	3	4	13	67	46	0.1667	0.2353	0.1951	0.5875	17	4	20	13	43
15	0.0000	0.0000	0.0000	0.7931	11	0	7	11	69	45	0.0000	0.0000	0.0000	0.7625	19	0	0	19	61
14	0.0000	0.0000	0.0000	0.8046	12	0	5	12	70	44	0.5000	0.2000	0.2857	0.9375	5	1	1	4	74
13	0.0588	1.0000	0.1111	0.8161	1	1	16	0	70	43	0.2500	1.0000	0.4000	0.9625	1	1	3	0	76
12	0.0000	0.0000	0.0000	0.9310	6	0	0	6	81	42	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
11	0.0000	0.0000	0.0000	0.9195	7	0	0	7	80	41	0.0000	0.0000	0.0000	0.9625	3	0	0	3	77
Left Quadrant (Q2 vs Q3)																			
21	0.0000	0.0000	0.0000	0.9425	5	0	0	5	82	31	0.0000	0.0000	0.0000	0.9750	2	0	0	2	78
22	0.0000	0.0000	0.0000	0.9540	4	0	0	4	83	32	0.0000	0.0000	0.0000	0.9875	1	0	0	1	79
23	0.0000	0.0000	0.0000	0.8736	3	0	8	3	76	33	0.0000	0.0000	0.0000	0.9625	1	0	2	1	77
24	0.0000	0.0000	0.0000	0.8621	12	0	0	12	75	34	0.0000	0.0000	0.0000	0.9500	4	0	0	4	76
25	0.2000	0.0769	0.1111	0.8161	13	1	4	12	70	35	0.0000	0.0000	0.0000	0.7375	9	0	12	9	59
26	0.5556	0.2778	0.3704	0.8046	18	5	4	13	65	36	0.3000	0.6316	0.4068	0.5625	19	12	28	7	33
27	0.5238	0.7097	0.6027	0.6667	31	22	20	9	36	37	0.4286	0.2609	0.3243	0.6875	23	6	8	17	49
28	0.8506	1.0000	0.9193	0.8506	74	74	13	0	0	38	0.8125	1.0000	0.8966	0.8125	65	65	15	0	0

Table 4.24: **3D 16+1 Dynamit + Hybrid Augmentation Test Results: Per-Tooth Analysis.** Comparing Upper Jaw (Maxillary) vs. Lower Jaw (Mandibular). *Supp* indicates Support.

Upper Jaw (Maxillary)										Lower Jaw (Mandibular)									
FDI	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**	**FDI**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Supp.**	**TP**	**FP**	**FN**	**TN**
Right Quadrant (Q1 vs Q4)																			
18	0.8256	0.9861	0.8987	0.8161	72	71	15	1	0	48	0.7436	0.9831	0.8467	0.7375	59	58	20	1	1
17	0.5152	0.5667	0.5397	0.6667	30	17	16	13	41	47	0.3871	0.7500	0.5106	0.7125	16	12	19	4	45
16	0.3750	0.5625	0.4500	0.7471	16	9	15	7	56	46	0.5000	0.2353	0.3200	0.7875	17	4	4	13	59
15	0.2500	0.1818	0.2105	0.8276	11	2	6	9	70	45	1.0000	0.1053	0.1905	0.7875	19	2	0	17	61
14	0.2308	0.2500	0.2400	0.7816	12	3	10	9	65	44	0.1176	0.4000	0.1818	0.7750	5	2	15	3	60
13	0.0588	1.0000	0.1111	0.8161	1	1	16	0	70	43	0.0000	0.0000	0.0000	0.8625	1	0	10	1	69
12	0.1875	0.5000	0.2727	0.8161	6	3	13	3	68	42	0.0000	0.0000	0.0000	0.9125	3	0	4	3	73
11	0.2500	0.8571	0.3871	0.7816	7	6	18	1	62	41	0.5000	0.3333	0.4000	0.9625	3	1	1	2	76
Left Quadrant (Q2 vs Q3)																			
21	0.1600	0.8000	0.2667	0.7471	5	4	21	1	61	31	0.1667	0.5000	0.2500	0.9250	2	1	5	1	73
22	0.1905	1.0000	0.3200	0.8046	4	4	17	0	66	32	0.0833	1.0000	0.1538	0.8625	1	1	11	0	68
23	0.0370	0.3333	0.0667	0.6782	3	1	26	2	58	33	0.1667	1.0000	0.2857	0.9375	1	1	5	0	74
24	0.4615	0.5000	0.4800	0.8506	12	6	7	6	68	34	0.2500	0.2500	0.2500	0.9250	4	1	3	3	73
25	0.5455	0.4615	0.5000	0.8621	13	6	5	7	69	35	0.3333	0.2222	0.2667	0.8625	9	2	4	7	67
26	0.4000	0.3333	0.3636	0.7586	18	6	9	12	60	36	0.8000	0.2105	0.3333	0.8000	19	4	1	15	60
27	0.5581	0.7742	0.6486	0.7011	31	24	19	7	37	37	0.2982	0.7391	0.4250	0.4250	23	17	40	6	17
28	0.8488	0.9865	0.9125	0.8391	74	73	13	1	0	38	0.8205	0.9846	0.8951	0.8125	65	64	14	1	1

Chapter 5

Discussion

The main limitation in our experiments is the scarcity of missing anterior teeth, which leads to an extreme class imbalance. Although our augmentation increases exposure to anterior-missing patterns during training, the effective number of true anterior-missing cases in the filtered test set remains very small (167 cases in total after id-matching), making anterior-region evaluation unstable. As a result, jaw-type prediction can remain relatively strong in several settings while tooth-level detection in the anterior region remains unreliable due to very limited true positives and reduced fine-grained visual evidence in top-view renderings.

5.1 Architectural Selection: Why We Discarded the 32-Neuron Approach

Before analyzing the performance of our proposed 16+1 architecture, it is necessary to address the preliminary experiments conducted with a global 32-neuron formulation. In this initial setup, the model was tasked with predicting the status of the entire dentition (32 permanent teeth) simultaneously from a single-jaw input scan.

Superficially, this architecture demonstrated robust convergence and superior metrics. As summarized in Table 5.1, the 2D Augmented + Dynamit configuration achieved a Macro-F1 score of **0.9110** and a Recall of **0.9264** on the test set. Numbers that ostensibly outperform the 16+1 results reported in Chapter 4.

Table 5.1: **2D 32-Neuron Architecture Performance.** Comparison of Macro-F1 and Recall scores. Although the metrics appear high (Test Recall > 0.92), we argue that this performance is inflated by structural label shortcuts rather than discriminative visual learning.

Configuration	Training Set		Test Set	
	Macro Recall	Macro F1	Macro Recall	Macro F1
Baseline (BCE)	0.9716	0.9767	0.8748	0.8892
Dynamit Loss	0.9763	0.9799	0.8984	0.8784
Augmented + Dynamit	0.8918	0.9359	**0.9264**	**0.9110**

5.1.1 The Illusion of High Performance: Structural Shortcuts

We argue that these high metrics are misleading and primarily driven by a **structural bias** in the problem formulation rather than effective tooth detection capabilities.

The 32-neuron vector compels the network to predict the status of both jaws, yet the input data (intraoral scan) captures only a single jaw (maxillary or mandibular). Consequently, for any given input, 16 of the 32 output labels correspond to the opposing, non-visible jaw. In our data preparation and evaluation pipeline, teeth belonging to the non-visible opposing jaw are treated as absent during supervision.

This creates a trivial shortcut for the network:

1. **Jaw Classification as a Proxy:** The model first learns to classify the jaw type (Upper vs. Lower), which is a relatively simple global shape recognition task.

2. **Deterministic Filling:** Once the jaw is identified, the network can blindly predict "missing" for the entire opposing arch (16 neurons) with 100% accuracy without examining any local features.

This phenomenon inflates global accuracy and recall metrics because half of the prediction task does not require visual evidence of tooth absence. The network effectively "cheats" by exploiting the deterministic label correlation between jaw type and the opposing dentition.

5.1.2 Justification for Exclusion

Because of this confounding variable, the 32-neuron results do not faithfully reflect the model's ability to detect missing teeth based on visual or geometric evidence. The apparent success masks potential failures in detecting truly ambiguous cases within the visible jaw.

Therefore, to ensure that the model is truly learning from visual features rather than guessing based on shortcuts, we excluded the 32-neuron results from our main analysis. We instead adopted the 16+1 architecture, which separates jaw classification from tooth detection. This forces the model to focus only on the visible teeth, removing the bias and ensuring that our reported metrics reflect true diagnostic performance.

5.2 2D Results: Architectural Trade-offs and Generalization

This section presents a comparative analysis of the 32-neuron and 16+1-neuron architectures across the training and Testing phases. We evaluate three configurations for 32-neuron model: Baseline (BCE Loss), Dynamit Loss, and Augmented Dataset + Dynamit Loss. For 16+1-neuron setting, we evaluate three configurations: Baseline (BCE Loss), Dynamit (Original Data), Dynamit + Hybrid Augmentation (Orig + Target + Random).

5.2.1 2D 16+1-Neuron Architecture: Generalization Challenges in Anterior Regions

The 16+1 architecture processes jaws independently, attempting to mitigate the complexity of full-arch analysis. However, our results reveal that this 2D rendering-based approach suffers from a severe generalization gap, particularly when addressing the extreme class imbalance inherent in anterior tooth detection.

The Generalization Gap: Overfitting to Augmented Patterns

A direct comparison between training and testing performance reveals a stark disparity. As detailed in Table 5.2, the model achieves near-perfect F1 scores (> 0.98) on the training set for almost all tooth positions, suggesting that the Hybrid Augmentation strategy successfully "forced" the model to learn rare patterns during training.

However, this performance collapses on the test set. For instance, Tooth 21 (Upper Left Central Incisor) drops from a perfect **1.0** in training to **0.0** in testing. We attribute this failure to a domain gap in the rendering pipeline: training samples are generated from a canonical top-down view, whereas test samples involve variable poses and scan qualities. The model appears to have overfitted to the specific artifacts of the synthetic training data rather than learning robust, view-invariant anatomical features.

Table 5.2: **Generalization Gap: Training vs. Testing Performance (16+1 Augmented).** Comparison of F1-scores. While the near-perfect training scores under Hybrid augmentation suggest that the model fits the synthetic rendering cues well, but this does not transfer to the real test domain, indicating overfitting to synthetic features.

FDI Tooth	Training F1	Testing F1	Gap (Δ)
Anterior Region (Incisors)			
Tooth 11 (Upper Central)	0.9922	0.1429	-0.8493
Tooth 21 (Upper Central)	0.9905	0.0000	-0.9905
Tooth 41 (Lower Central)	0.9583	0.0000	-0.9583
Posterior Region (Molars)			
Tooth 18 (Upper Wisdom)	0.9942	0.8844	-0.1098
Tooth 48 (Lower Wisdom)	0.9849	0.8710	-0.1139

The Precision-Recall Trade-off under Dynamit Loss

The impact of the Dynamit loss function becomes evident when analyzing the trade-off between identifying missing teeth (Recall) and minimizing false alarms (Precision).

In the baseline BCE model, the overwhelming prevalence of "present" teeth encourages the network to default to predicting "present" to minimize global error. This results in a high Accuracy but poor sensitivity to missing teeth. By introducing the Dynamit loss and Hybrid Augmentation, we explicitly penalized this behavior.

As shown in our Confusion Matrix analysis (Table 4.19), the Dynamit + Hybrid Aug model correctly identified **356** missing teeth (True Positives), significantly more than the Baseline's **315**.

However, this sensitivity comes at a cost: False Positives rose from 89 (Baseline) to **275** (Dynamit + Aug). This shift indicates that while the Dynamit loss successfully counteracts the class imbalance by forcing the model to attend to "missing" signals, the 2D visual evidence is often too ambiguous to distinguish true gaps from rendering artifacts, leading to over-prediction of missing teeth. This is consistent with augmentation increasing the prevalence and diversity of missing patterns during training, which shifts the model toward higher sensitivity but also raises false-alarm risk under test-domain ambiguity.

Region-Specific Failure: The "Anterior Collapse"

The failure of the 2D 16+1 architecture is geographically specific. We term this phenomenon the **"Anterior Collapse"**. As summarized in Table 5.3, the model maintains respectable performance on posterior teeth (e.g., wisdom teeth 18 and 48), yet collapses to near-zero performance on incisors (notably 21/31/41), with 11 remaining very low.

Table 5.3: **The Anterior Collapse: Performance Disparity by Anatomical Region.** Test set results for the 16+1 Dynamit Augmented model. The model effectively detects posterior missing teeth (larger visual surface) but fails to detect anterior missing teeth (smaller visual footprint).

Region	Tooth (FDI)	Precision	Recall	F1-Score
Posterior (Molars)	18 (Upper Right)	0.8667	0.9028	0.8844
	28 (Upper Left)	0.8947	0.9067	0.9007
	38 (Lower Left)	0.9531	0.9125	0.9324
	48 (Lower Right)	0.8308	0.9153	0.8710
Anterior (Incisors)	11 (Upper Central)	0.1429	0.1429	0.1429
	21 (Upper Central)	0.0000	0.0000	0.0000
	31 (Lower Central)	0.0000	0.0000	0.0000
	41 (Lower Central)	0.0000	0.0000	0.0000

This disparity highlights a fundamental limitation of the 2D rendering approach. Molar teeth are bulky, creating large, distinct "holes" in a 2D depth map when missing. Conversely, incisors are thin. When an incisor is missing, the gap is narrow and easily obscured by neighboring teeth or rendering noise. The 16+1 architecture, lacking the 3D geometric context of the full arch, cannot reliably differentiate these subtle visual cues from noise, leading to the observed collapse in anterior performance.

Global Context vs. Local Feature Extraction

Finally, it is worth noting the contrast between the model's global and local capabilities. While local anterior detection failed, the auxiliary jaw classification neuron achieved an accuracy of **88.02%** on the test set (Table 4.21).

Table 5.4: **Global Context vs. Local Detection Capabilities.** The architecture succeeds at global categorization (Jaw Type) but struggles with fine-grained local detection in noisy regions (Anterior Teeth).

Task Type	Test Set Performance
Global Context	
Jaw Classification Accuracy (Auxiliary Neuron)	**88.02%**
Local Detection	
Posterior Tooth Detection (Avg. F1, e.g., 38)	$\approx 93.00\%$
Anterior Tooth Detection (Avg. F1, e.g., 41)	**0.00%**

This result confirms that the network can learn global anatomical features (distinguishing Maxilla from Mandible) even from imperfect 2D renderings. However, the resolution and information loss inherent in projecting 3D scans to 2D images prove critical when attempting to resolve fine-grained features like anterior tooth absence.

5.3 3D Results Analysis

5.3.1 3D 16+1-Neuron Architecture: Resolving the Anterior Generalization Gap

Unlike the 2D rendering-based approach, the 3D point cloud pathway processes the intrinsic geometric structure of the dental arch. Our results demonstrate that this modality shift, combined with Hybrid Augmentation, effectively resolves the "Anterior Collapse" observed in 2D models.

Structural Robustness Overcomes Viewpoint Dependency

The most significant finding in the 3D experiments is the restoration of performance in the anterior region. While the 2D augmented model failed to generalize on incisors (Test F1 = 0.00 for Tooth 21), the 3D Augmented + Dynamit model achieved a Recall of **0.9765** and an F1-score of **0.6721** for the same tooth.

We attribute this success to the view-invariant nature of 3D point clouds. In 2D, the visual evidence of a missing anterior tooth is a narrow gap that can be easily occluded or mistaken for shadowing artifacts depending on the rendering angle. In contrast, the 3D architecture (PointNet++) operates directly on spatial coordinates and surface normals. A missing incisor creates a distinct topological depression in the gumline, a geometric feature that remains consistent regardless of the scan's orientation. This allows the model to learn robust representations of "absence" that generalize well to unseen test data, substantially mitigating the domain gap that limits the 2D approach.

Breaking the Zero-Shot Barrier with Hybrid Augmentation

The comparison between the 3D Baseline and the 3D Augmented models further validates the necessity of targeted synthetic data.

As shown in Table 4.42, the Baseline (BCE) model exhibits the same "zero-shot" failure mode

seen in 2D: anterior teeth such as 11, 21, 31, and 41 have a Recall of **0.00**, confirming that the natural scarcity of these samples in the training set prevents standard models from learning discriminative features.

However, the introduction of Hybrid Augmentation completely reverses this trend. In the Augmented model (Table 4.33), Recall for these same teeth jumps to near-perfect levels (e.g., Tooth 11 Recall: **0.9885**, Tooth 31 Recall: **0.9565**). This confirms that the synthetic generation of missing teeth provided the necessary support for the network to define decision boundaries for these minority classes, effectively transforming a zero-shot problem into a supervised learning task.

The Sensitivity-Specificity Trade-off

While the 3D Augmented model excels at detecting missing teeth (High Recall), it exhibits a trade-off in Precision. For example, Tooth 11 achieves a Recall of 0.9885 but a Precision of **0.5276**.

This behavior is a direct consequence of the Dynamit loss function interacting with the extreme class imbalance. The Dynamit loss is designed to aggressively penalize missed detections (False Negatives) to combat the dominance of the "present" class. In the test set, where actual missing anterior teeth are rare, even a small number of False Positives (predicting a gap where there is tight crowding or spacing) significantly dilutes the Precision score.

Table 4.40 supports this observation: the Dynamit + Hybrid Aug model achieves the highest overall Macro Recall (**0.5565**) and Tooth Balanced Accuracy (**0.7707**) among all configurations, but its Macro Precision (**0.3769**) is lower than its Recall. This indicates that the model has adopted a "high-sensitivity" operating point, clinically preferable for a screening tool where missing a diagnosis (False Negative) is more detrimental than a false alarm.

Global Context Consistency

Similar to the 2D results, the 3D architecture maintains high accuracy in global context recognition. The auxiliary jaw classification neuron achieves an accuracy of **89.82%** on the test set. This suggests that 3D point clouds, like 2D images, contain sufficient global structural cues (e.g., arch curvature, palate shape) to distinguish maxillary from mandibular scans effectively, independent of the local tooth-wise predictions.

5.4 Comparative Analysis: 2D vs. 3D Pathways

This section provides a systematic comparison of the 2D render-based and 3D point cloud-based pathways under controlled experimental conditions. We evaluate both approaches across three critical dimensions: (i) per-tooth detection performance, particularly on rare anterior missing teeth, (ii) training efficiency and convergence behavior, and (iii) computational requirements for clinical deployment. Our analysis reveals fundamental differences in how each modality handles the extreme class imbalance inherent in tooth presence classification.

5.4.1 Test-Set Performance: The Decisive Role of Input Modality

Quantitative Summary

Table 5.5 presents an aggregate comparison of the best-performing configurations from each pathway on the held-out test set. The 3D pathway demonstrates a substantial advantage across all metrics, particularly in Balanced Accuracy, the most clinically relevant metric given the severe class imbalance.

Table 5.5: Test-Set Performance Comparison: Best Configuration from Each Pathway

Pathway	Config	Prec.	Rec.	F1	Acc.	Bal. Acc.
2D Render	Dynamit + Hybrid Aug	0.3092	0.3101	0.2873	0.8220	0.5713
3D Point Cloud	Dynamit + Hybrid Aug	0.3769	0.5565	0.3743	0.7994	**0.7707**
Absolute Improvement (3D over 2D)		+0.0677	+0.2464	+0.0870	-0.0226	**+0.1994**
Relative Improvement (%)		+21.9%	+79.5%	+30.3%	-2.7%	**+34.9%**

Key Observation: The 3D pathway achieves a **+19.94 percentage point** improvement in Balanced Accuracy despite slightly lower standard accuracy. This divergence confirms that standard accuracy is dominated by the trivial task of correctly predicting present teeth (the majority class), whereas Balanced Accuracy properly weights performance on the clinically critical minority class (missing teeth). The 79.5% relative improvement in Recall indicates that the 3D model detects substantially more true missing teeth, though at a modest precision cost.

The "Anterior Collapse" Phenomenon in 2D

A region-specific analysis reveals a systematic failure mode in the 2D pathway: **anterior teeth (incisors and canines) exhibit near-zero detection capability**, a pattern we term the *Anterior Collapse*. Table 5.6 compares per-tooth F1-scores for representative anterior positions.

Table 5.6: Per-Tooth F1-Scores: Anterior Region Performance Breakdown

Tooth (FDI)	Anatomical Name	2D Pathway		3D Pathway	
		Support	F1	Support	F1
11	Upper Right Central Incisor	7	0.1429	7	0.3871
21	Upper Left Central Incisor	5	0.0000	5	0.2667
31	Lower Left Central Incisor	2	0.0000	2	0.2500
41	Lower Right Central Incisor	3	0.0000	3	0.4000
13	Upper Right Canine	1	0.0000	1	0.1111
23	Upper Left Canine	3	0.0000	3	0.0667
Anterior Mean (6 teeth)		–	**0.0238**	–	**0.2469**

Diagnostic Interpretation: The 2D model's failure on anterior teeth stems from two converging factors:

1. **Insufficient Visual Evidence in Top-View Renderings:** Anterior teeth occupy a much smaller image area compared to molars in the canonical top-view projections used for training. Missing anterior teeth manifest as narrow gaps that are visually indistinguishable from normal interproximal spacing or rendering artifacts (e.g., specular highlights, mesh repair noise). In contrast, missing molars create large, unambiguous voids in the posterior region.

2. **Extreme Training Scarcity Amplified by Projection:** Even with augmentation, anterior missing teeth remain the rarest class (Support < 7 per position in the test set). The 2D projection discards geometric context, such as alveolar ridge curvature and gingival topology, that would provide auxiliary cues for absence detection. The network thus relies solely on pixel-level gap detection, a task for which it has seen a limited number of effective examples during training.

3D Geometric Advantage: The 3D pathway mitigates both issues by processing raw surface geometry. Missing teeth create a *topological depression* in the point cloud, a concave region where a convex tooth crown is expected. This geometric signature is:

- **View-invariant:** Independent of camera angle or projection artifacts.
- **Robust to noise:** Local surface curvature features (captured by PointNet's neighborhood aggregation) persist despite mesh repair artifacts.
- **Contextual:** The point cloud preserves spatial relationships between adjacent teeth and gingival surfaces, enabling the network to infer absence from surrounding structures.

For Tooth 21 (Upper Left Central Incisor), the 3D model achieves Precision = 0.1600, Recall = 0.8000, F1 = 0.2667. While this F1-score remains low in absolute terms, it represents a **qualitative breakthrough**: the model successfully detects 4 out of 5 missing instances, compared to 0 out of 5 in the 2D baseline. This confirms that 3D geometry provides the minimum information content required for anterior tooth absence detection under realistic data constraints.

Posterior Teeth: Convergence of Performance

In contrast to the anterior collapse, both pathways perform competently on posterior teeth (molars), particularly wisdom teeth (FDI 18, 28, 38, 48). Table 5.7 summarizes results for these high-absence-rate positions.

Table 5.7: Posterior Teeth Performance: Wisdom Teeth (High Support)

Tooth (FDI)	Pathway	Support	Recall	F1
18 (Upper Right)	2D	72	0.9028	0.8844
	3D	72	0.9861	0.8987
28 (Upper Left)	2D	74	0.9189	0.9067
	3D	74	0.9865	0.9125
38 (Lower Left)	2D	65	0.9385	0.9457
	3D	65	0.9846	0.8951
48 (Lower Right)	2D	59	0.9153	0.8710
	3D	59	0.9831	0.8467
Mean Recall (4 wisdom teeth)		–	**2D: 0.9189**	**3D: 0.9886**

Analysis: Both pathways achieve >90% recall on wisdom teeth, where large anatomical gaps and high training support (≥59 samples) provide sufficient learning signal. The 3D pathway maintains a slight edge (+6.97 percentage points in mean recall), attributed to rotation invariance: PointNet's PCA-based normalization ensures consistent feature extraction regardless of scan orientation, whereas 2D renderings from unconstrained test-set orientations introduce viewpoint variability.

5.4.2 Training Dynamics: Convergence and Augmentation Efficacy

Beyond test-set performance, the two pathways exhibit markedly different training behaviors, particularly in their response to data augmentation.

Training Set Convergence

Table 5.8 compares training-set metrics under the Dynamit + Hybrid Augmentation configuration. Both pathways achieve high training performance, but the 3D pathway converges to slightly lower metrics, a counterintuitive result that warrants explanation.

Table 5.8: Training-Set Performance: Dynamit + Hybrid Augmentation

Pathway	Precision	Recall	F1-Score	Accuracy
2D Render	0.9428	0.9554	0.9487	0.9754
3D Point Cloud	0.8915	0.9107	0.9006	0.9520
Difference	-0.0513	-0.0447	-0.0481	-0.0234

Interpretation: The 2D pathway's higher training metrics are partially attributable to **task simplification via fixed viewpoints**. All training images are rendered from a canonical top-view orientation, creating a visually consistent visual domain that facilitates memorization. The network learns to associate specific pixel patterns without needing to generalize across viewpoints.

In contrast, the 3D pathway processes raw point clouds with variable densities, surface noise, and PCA-induced rotations. Even after normalization, point cloud inputs exhibit higher intrinsic variability than controlled 2D renders, making the training task inherently harder. The lower training F1 (0.9006 vs. 0.9487) reflects this increased difficulty, but paradoxically, this difficulty enforces learning of *more robust, view-invariant features* that transfer better to the unconstrained test set.

Evidence from Test-Set Generalization Gap: We quantify the generalization gap as the difference between training and test F1-scores:

$$\text{Generalization Gap}_{2D} = 0.9487 - 0.2873 = \mathbf{0.6614}$$
$$\text{Generalization Gap}_{3D} = 0.9006 - 0.3743 = \mathbf{0.5263}$$

The 2D pathway suffers a 66.14 percentage point drop, compared to 52.63 points for 3D. This confirms that the 2D model's high training performance is partially due to overfitting to canonical viewpoints, whereas the 3D model learns features that generalize more consistently.

Impact of Augmentation: Breaking the Zero-Shot Barrier

Both pathways initially exhibit near-zero recall on anterior teeth when trained solely on real data (Baseline BCE). Table 5.9 quantifies the improvement brought by Hybrid Augmentation.

Table 5.9: Effect of Augmentation on Macro-Averaged Recall (Test Set)

Pathway	Baseline (BCE)	Dynamit + Hybrid Aug	Δ Recall
2D Render	0.1873	0.3101	**+0.1228**
3D Point Cloud	0.2047	0.5565	**+0.3518**

Key Insight: Augmentation provides a **0.3518 relative improvement** in 3D recall compared to 2D. This disparity arises because:

1. **Synthetic Fidelity:** The 3D augmentation pipeline (Delete-and-Fill) operates directly on mesh geometry, producing synthetic missing teeth that are geometrically indistinguishable from real absences, both manifest as healed gingival surfaces. In contrast, 2D synthetic renders inherit the viewpoint consistency of the training set, failing to expose the network to the full range of orientations present in test images.

2. **Feature Transferability:** Geometric absence features (concave depressions, alveolar ridge gaps) learned from synthetic 3D data transfer seamlessly to real test scans. Visual absence features learned from synthetic 2D renders (specific gap patterns at specific pixel locations) do not transfer to test images rendered from different angles or with different lighting.

5.4.3 Auxiliary Task Performance: Jaw Classification

Both architectures include a +1 auxiliary neuron to classify jaw type (Upper vs. Lower). Table 5.10 reports test-set accuracy for this task.

Table 5.10: Jaw Classification Accuracy (Test Set, +1 Neuron)

Configuration	2D Pathway	3D Pathway
Baseline (BCE)	0.7545	0.9341
Dynamit (Original Data)	0.9042	0.9281
Dynamit + Hybrid Aug	0.8802	0.8982

- **3D Baseline Superiority:** The 3D baseline achieves 93.41% jaw accuracy, compared to 75.45% in 2D. This suggests that global anatomical features (e.g., dental arch curvature, which differs between maxilla and mandible) are more robustly encoded in 3D geometry than in 2D projections.
- **Augmentation Impact:** Interestingly, augmentation *degrades* jaw classification accuracy in both pathways. This is likely due to **task interference**: the aggressive augmentation strategies (removing multiple teeth, altering gingival topology) may obscure global arch shape cues used for jaw classification, forcing the network to allocate representational capacity to the primary tooth-level task.
- **Clinical Relevance:** Jaw classification accuracy >88% is acceptable for clinical use, as scans are typically labeled by acquisition protocol (operators know which jaw they are scanning). The auxiliary neuron primarily serves as a *sanity check* to detect mislabeled data rather than as a critical prediction target.

5.4.4 Synthesis: When Does 3D Outperform 2D?

Our results provide empirical evidence for a principled answer to the representation debate. The 3D point cloud pathway achieves superior performance when three conditions converge:

1. **Task Requires Fine-Grained Spatial Reasoning:** Tooth absence detection relies on recognizing subtle geometric deviations (gaps, concavities) that are view-dependent in 2D but intrinsic to 3D structure.
2. **Training Data Exhibits Extreme Class Imbalance:** With <5% positive samples for anterior teeth, the 3D pathway's geometric features provide a more sample-efficient learning signal than 2D pixel patterns.
3. **Test Distribution Differs from Training:** The test set contains scans from different scanners, orientations, and mesh quality levels. 3D features (curvature, topology) generalize across these variations, whereas 2D features (pixel intensities, edge patterns) are domain-specific.

Conversely, 2D approaches remain viable for tasks with:

- High-level semantic classification (e.g., "presence of restoration" vs. "no restoration")
- Abundant training data across all classes
- Controlled imaging conditions (consistent viewpoints, lighting, resolution)

For automated dental charting under realistic clinical constraints, our findings strongly favor the 3D pathway.

5.4.5 Limitations of Current Comparison

Our study compares specific instantiations of 2D and 3D pipelines. The following factors may influence generalizability:

- **2D Architecture:** We used ResNet-18 with ImageNet initialization. Larger models (e.g., ResNet-50, EfficientNet-B7) or domain-adaptive pre-training (e.g., self-supervised learning on dental images) might reduce the performance gap.
- **3D Architecture:** PointNet is a relatively simple 3D architecture. More sophisticated models (e.g., PointNet++, DGCNN, Point Transformer) may achieve further improvements, though at higher computational cost.
- **Multi-View Fusion:** Our 2D pipeline explored multi-view rendering with confidence-based selection during early experiments. Alternative fusion strategies (e.g., attention mechanisms, learned view selection) could potentially improve 2D generalization.

5.4.6 Summary of Key Findings

Our comparative analysis establishes three primary conclusions:

1. **3D Superiority for Rare Class Detection:** The 3D point cloud pathway achieves +34.9% relative improvement in Balanced Accuracy, driven by robust detection of anterior missing teeth, a task where 2D approaches systematically fail.
2. **Augmentation is Necessary but Modality-Dependent:** Hybrid augmentation provides 2.7$\times$ greater recall improvement in 3D than in 2D, indicating that synthetic fidelity is modality-specific. Geometric augmentation transfers better than visual augmentation.
3. **2D Limitations Under Unconstrained Acquisition:** Fixed-view 2D renderings struggle under real-world variability in scan orientation and quality, particularly for anatomically small structures such as incisors.

These findings provide actionable guidance for researchers and practitioners developing automated dental analysis systems: when working with 3D scan data, process it in its native geometric representation rather than projecting to 2D, especially for tasks requiring fine-grained spatial reasoning under data scarcity.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contributions

This project investigated automated tooth presence classification in intraoral scans under a small-data, highly imbalanced setting. Through a systematic comparison between 2D render-based and 3D point cloud-based pathways, this work clarifies the strengths and limitations of each modality under our evaluation protocol. The primary contributions are:

- **Architectural Design (16+1 Strategy):** We introduced a jaw-specific "16+1" neuronal architecture that decouples the detection of individual teeth from global jaw classification. This formulation removes structural shortcuts inherent in global labeling, enforcing the learning of fine-grained anatomical features.
- **Data-Centric Solution for Zero-Shot Learning:** We developed a "Hybrid Augmentation" pipeline that combines real data with targeted synthetic deletion. This strategy successfully transformed the "zero-shot" problem of anterior tooth detection, where real training samples were virtually non-existent, into a solvable supervised learning task.
- **Definitive Modality Comparison:** We provided empirical evidence resolving the 2D vs. 3D debate for this specific clinical application. Our results quantify the "Anterior Collapse" in 2D approaches and demonstrate the necessity of 3D geometric reasoning for detecting rare, subtle anatomical absences.

6.2 Primary Findings

The experimental results yield three fundamental conclusions regarding the automation of dental diagnostics:

6.2.1 The Advantage of 3D Geometry

Our comparative analysis indicates that, under our dataset and evaluation protocol, native 3D processing is the more effective modality for tooth presence classification. The 3D point cloud pathway (PointNet) achieved a +19.94% improvement in Balanced Accuracy over the best 2D configuration on the held-out test set. In particular, the 3D model showed improved sensitivity to missing anterior teeth by leveraging geometric cues around the gingival region that are largely degraded by 2D projection.

6.2.2 Anterior Collapse in 2D

While the 2D render-based approach was computationally efficient and performed reasonably on global cues (e.g., jaw classification) and high-support posterior teeth, it showed a systematic failure on missing anterior teeth. We observed an *Anterior Collapse*, where per-tooth recall and F1-scores for incisors were near zero. This behavior is likely influenced by the limited effective resolution of anterior gaps in canonical top-down views and the ambiguity between true absence cues and rendering or occlusion artifacts when depth geometry is not directly available.

6.2.3 The Critical Role of Distribution-Aware Training

Algorithmic complexity alone was insufficient to overcome the dataset's long-tail distribution. Standard Baseline models (BCE) in both 2D and 3D failed to detect minority classes. The integration of the **Dynamit loss function** combined with **Hybrid Augmentation** was the decisive factor in performance. This combination shifted the operating point to prioritize the minority "missing" class, improving recall by over 170% in the 3D pathway, confirming that targeted data synthesis is a prerequisite for medical AI in small-data regimes.

6.3 Limitations

Despite the advances presented, several limitations remain which define the scope of the current success:

- **Precision-Recall Trade-off:** While our methods significantly improved Recall (sensitivity), this came at the cost of Precision. The model currently exhibits a higher rate of False Positives, tending to flag ambiguous gaps (e.g., spacing or crowding) as missing teeth.
- **Static Training Viewpoints (2D):** The 2D pathway was trained on canonical top-down views. While this simplified training, it contributed to poor generalization on the unconstrained test set. A dynamic multi-view training pipeline was not implemented in this scope.
- **Basic 3D Backbone:** We utilized PointNet, a pioneering but relatively simple architecture. We did not explore local-feature-aware networks (like DGCNN) which might better resolve the fine-grained boundaries between gums and teeth.

6.4 Future Work

To address the limitations identified above and further bridge the gap between algorithmic research and clinical deployment, we propose the following directions for future investigation:

1. **Exploration of Advanced Backbones and Pre-training:** While this study established baselines using ResNet and PointNet, recent advancements in foundational models warrant exploration.
    - **For 2D:** Transitioning from standard CNNs to Vision Transformers (ViT) or networks pre-trained specifically on medical imaging datasets could enhance feature extraction capabilities for noisy dental renderings.

- **For 3D:** Moving beyond PointNet to local-feature-aware architectures like DGCNN or advacnced point Transformers may better capture fine-grained gingival topology. Furthermore, leveraging self-supervised pre-training on larger collections of unlabeled intraoral scans, if available.

2. **Robust Multi-View Training for 2D:** Our analysis revealed that the 2D pathway suffered from overfitting to the canonical top-down view used during training. Future work should implement a *dynamic multi-view training strategy*, where the network is exposed to randomized rendering angles (varying azimuth and elevation) during the training phase itself. This would enforce the learning of view-invariant representations, theoretically narrowing the generalization gap observed on the unconstrained test set.

3. **Clinical Integration and Interactive Visualization:** To translate algorithmic success into clinical utility, developing a user-friendly **visualization application** is a critical next step. Such an interface should:

   - Visualize detection results directly on the 3D intraoral mesh rather than as abstract labels.
   - Provide a "human-in-the-loop" mechanism, allowing clinicians to verify or correct predictions, which can subsequently be used to fine-tune the model.
   - Display confidence heatmaps to intuitively communicate model uncertainty to the practitioner.

4. **Active Learning for Rare Patterns:** Given that anterior tooth loss remains a statistically rare event ($< 5\%$ prevalence), relying solely on random data collection is inefficient. An active learning framework could automatically identify and prioritize scans with ambiguous or suspected anterior absences for expert annotation. This would maximize sample efficiency and rapidly improve performance on the "long tail" of the data distribution.

5. **Uncertainty Quantification for Trustworthy AI:** The precision-recall trade-offs observed in our experiments highlight the risk of false positives. Incorporating uncertainty estimation techniques (e.g., Monte Carlo Dropout or Evidential Deep Learning) would allow the system to flag low-confidence predictions. In a clinical context, the system could refrain from making an automated decision in uncertain cases and instead prompt the dentist for manual review, thereby enhancing clinical trust and safety.

6.5 Closing Remarks

This project aimed to address a pervasive challenge in digital dentistry: the automated and reliable detection of tooth presence in the face of extreme data scarcity and class imbalance. By introducing the "16+1" neuronal architecture, the Dynamit loss function, and a targeted Hybrid Augmentation strategy, we have established an experimental framework and a set of practical components that improve performance for automated tooth presence classification under severe data scarcity and class imbalance.

Our comparative investigation clarifies the modality trade-offs for this task under our evaluation protocol. While 2D rendering-based approaches are conceptually simpler, they suffer from an inherent anterior collapse due to the loss of geometric information in projection. In contrast, our findings demonstrate that the 3D point cloud pathway, when empowered by distribution-aware synthetic data, captures the intrinsic topological signatures of tooth absence. This confirms

that for fine-grained anatomical reasoning, preserving the native dimensionality of the data is paramount.

Ultimately, this work extends beyond the specific application of tooth detection. It serves as a case study for medical AI in the "small data" regime, illustrating that algorithmic innovations alone are often insufficient to overcome structural data deficits. Instead, the synergy of domain-specific data synthesis and geometry-aware deep learning offers a scalable path forward. We hope that the proposed methodology can serve as a foundation for future clinical-facing systems that reduce manual charting effort and support dental workflows.

6.5.1 Code Availability

To support reproducibility and facilitate future research, the complete implementation of the 2D and 3D pipelines, including data preprocessing, rendering scripts, training configurations, and evaluation code, is publicly available at:

https://github.com/looooopiii/Tooth-Presence-Classification/tree/main

List of Figures

List of Tables

Bibliography

Afrashthefar, K. I., Alnakeb, N. A., and Assery, M. K. M. (2022). Accuracy of intraoral scanners versus traditional impressions: A rapid umbrella review. *Journal of Evidence-Based Dental Practice*, 22(3):101719.

Al-Sarem, M., Al-Asali, M., Alqutaibi, A. Y., et al. (2022). Enhanced Tooth Region Detection Using Pretrained Deep Learning Models. *International Journal of Environmental Research and Public Health*, 19(22):15414.

Ben-Hamadou, A., Smaoui, O., Chaabouni-Chouayakh, H., Rekik, A., Pujades, S., Boyer, E., Strippoli, J., Thollot, A., Setbon, H., Trosset, C., et al. (2022). Teeth3ds: a benchmark for teeth segmentation and labeling from intra-oral 3d scans. *arXiv preprint arXiv:2210.06094*.

Ben-Hamadou, A., Smaoui, O., Rekik, A., et al. (2023a). 3DTeethSeg-22: 3D Teeth Scan Segmentation and Labeling Challenge. *arXiv preprint arXiv:2305.18277*. MICCAI Challenge.

Ben-Hamadou, A., Smaoui, O., Rekik, A., Pujades, S., Boyer, E., Lim, H., Kim, M., Lee, M., Chung, M., Shin, Y.-G., Leclercq, M., Cevidanes, L., Prieto, J. C., Zhuang, S., Wei, G., Cui, Z., Zhou, Y., Dascalu, T., Ibragimov, B., Yong, T.-H., Ahn, H.-G., Kim, W., Han, J.-H., Choi, B., van Nistelrooij, N., Kempers, S., Vinayahalingam, S., Strippoli, J., Thollot, A., Setbon, H., Trosset, C., and Ladroit, E. (2023b). 3dteethseg'22: 3d teeth scan segmentation and labeling challenge. *arXiv preprint arXiv:2305.18277*.

Chen, N. et al. (2023). Fully Automated Tooth Segmentation and Labeling for Both Full- and Partial-Arch Intraoral Scans Using Deep Learning. *Journal of Dentistry*, 138:104713.

Chung, M., Lee, J., Park, S., et al. (2021). Individual Tooth Detection and Identification from Dental Panoramic X-ray Images via Point-wise Localization and Distance Regularization. *Artificial Intelligence in Medicine*, 111:101996.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Ghamrawi, S. et al. (2024). TSegLab: Multi-stage 3D Dental Scan Segmentation and Labeling. *Computer Methods and Programs in Biomedicine*, 257:108466.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Impellizzeri, A., Horodynski, M., De Stefano, A., Palaia, G., Polimeni, A., Romeo, U., Guercio-Monaco, E., and Galluccio, G. (2020). Cbct and intra-oral scanner: The advantages of 3d

technologies in orthodontic treatment. *International Journal of Environmental Research and Public Health*, 17(24):9428.

Kim, E., Hwang, J. J., Cho, B.-H., et al. (2024). Classification of Presence of Missing Teeth in Each Quadrant Using Deep Learning Artificial Intelligence on Panoramic Radiographs of Pediatric Patients. *Journal of Clinical Pediatric Dentistry*, 48(3):80–88.

Lian, C., Wang, L., Wu, T.-H., Wang, F., Yap, P.-T., Ko, C.-C., and Shen, D. (2020). Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners. *IEEE Transactions on Medical Imaging*, 39(7):2440–2450.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal loss for dense object detection.

Park, J. M., Lee, J. S., Moon, S. G., and Lee, K. B. (2022). Deep Learning Based Detection of Missing Tooth Regions for Dental Implant Planning in Panoramic Radiographic Images. *Applied Sciences*, 12(3):1595.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2016). Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593.

Tan, Z. Q., Roscoe, M. G., Addison, O., and Li, Y. (2025). Deep learning in dentistry: A systematic review from an ai researcher viewpoint. *medRxiv*. Preprint, not peer-reviewed.

Truant, K. (n.d.). The International Tooth Numbering System (FDI). in *Oral & Maxillofacial Surgery – An Introduction for Registered Nurses*.

Wu, T., Lian, C., Lee, S., Pastewait, M., Piers, C., Liu, J., Wang, F., Wang, L., Jackson, C., Chao, W., Shen, D., and Ko, C. (2021). Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans. *CoRR*, abs/2109.11941.

Xie, S. and Tu, Z. (2015). Holistically-nested edge detection. *CoRR*, abs/1504.06375.

Xu, X., Liu, C., and Zheng, Y. (2021). Toward Clinically Applicable 3-Dimensional Tooth Segmentation via Deep Learning. *Journal of Dental Research*, 100(13):1449–1456.

Zhou, K. et al. (2023). Hierarchical Self-Supervised Learning for 3D Tooth Segmentation in Intra-Oral Mesh Scans. *IEEE Transactions on Medical Imaging*, 42(4):1066–1077.

Zhu, Y., Kechichian, R., Richert, R., Ikehata, S., and Valette, S. (2026). High-fidelity 3d tooth reconstruction by fusing intraoral scans and cbct data via a deep implicit representation.