Bits2Bites: Intra-oral Scans Occlusal Classification

Lorenzo Borghi*,1, Luca Lumetti*,1, Francesca Cremonini[2], Federico Rizzo[2], Costantino Grana[1], Luca Lombardo[2], and Federico Bolelli[1] ✉

[1] University of Modena and Reggio Emilia, Italy
[2] University of Ferrara, Italy

Abstract. We introduce *Bits2Bites*, the first publicly available dataset for occlusal classification from intra-oral scans, comprising 200 paired upper and lower dental arches annotated across multiple clinically relevant dimensions (sagittal, vertical, transverse, and midline relationships). Leveraging this resource, we propose a multi-task learning benchmark that jointly predicts five occlusal traits from raw 3D point clouds using state-of-the-art point-based neural architectures. Our approach includes extensive ablation studies assessing the benefits of multi-task learning against single-task baselines, as well as the impact of automatically-predicted anatomical landmarks as input features. Results demonstrate the feasibility of directly inferring comprehensive occlusion information from unstructured 3D data, achieving promising performance across all tasks. Our entire dataset, code, and pretrained models are publicly released to foster further research in automated orthodontic diagnosis.

Keywords: Intra-oral Scans, Dental Occlusion, 3D Point Cloud

1 Introduction

Deep learning has become a key enabler in dental healthcare, supporting the automation and enhancement of diagnostic workflows. The growing availability of public 3D imaging datasets related to dental healthcare has significantly contributed to the research community. For instance, in recent years, different datasets introduced labeled CBCT scans with dozens of anatomical structures, fostering research in segmenting complex regions such as the inferior alveolar canal, teeth, jaws, and dental implants [2,3,4,5,7,8,17]. In the domain of intra-oral 3D scanning (IOS), large-scale datasets like 3DTeethSeg [1] offer

Fig. 1: *Bits2Bites* logo.

† Equal contribution. Authors are allowed to list their name first on their CVs.
✉ Corresponding author: federico.bolelli@unimore.it.

full-tooth segmentation annotations across hundreds of scans, while the Teeth-Land dataset, released by the same authors, provides detailed landmark annotations for each tooth. These resources have catalyzed the development of a wide range of methods—from voxel-based and surface-based segmentation networks to point-based landmark detection approaches [10,13,14,16,19].

Despite these advances, several clinically relevant tasks remain underexplored in the context of 3D IOS analysis, largely due to the lack of publicly available annotations. One such task is *occlusal classification*, which involves determining the relationship between the upper and lower dentition when the mouth is closed. Accurate occlusion assessment is fundamental for orthodontic diagnosis and treatment planning, as it directly informs the strategy for interventions such as braces or clear aligners and serves as a baseline for evaluating treatment success. While prior work has investigated malocclusion detection from 2D snapshots of 3D models [11], these modalities lack the rich 3D surface information captured in IOS scans. Consequently, they omit crucial depth and structural cues that are essential for a fine-grained and comprehensive occlusion analysis.

To the best of our knowledge, no existing 3D deep learning method directly operates on paired upper and lower IOS meshes to predict occlusion classes. Addressing this gap, our work introduces new resources and benchmarks to facilitate progress in this direction.

Contribution. In summary, the contributions of this work are outlined below:

- We present *Bits2Bites*, the first publicly available dataset of 200 paired intra-oral scans with multi-dimensional clinical labels for occlusion classification, including sagittal, vertical, transverse, and midline relationships;[3]
- A robust multi-task and single-task learning benchmark is introduced for this task, evaluating two state-of-the-art point cloud backbones and demonstrating the effectiveness of jointly learning multiple occlusal traits;
- Detailed ablation studies are carried out to analyze the impact of using automatically-predicted anatomical landmarks as input features and to validate our multi-task learning strategy against single-task baselines;
- We release our entire codebase and pretrained models to ensure reproducibility and foster further research in the community.[4]

2 Dataset

The dataset comprises 200 pairs of registered intra-oral scans in STL format, with separate high-resolution meshes for the upper and lower dental arches. All scans are spatially aligned to preserve the true occlusal relationship between the jaws and are transformed to a shared reference RAS (Right-Anterior-Superior) frame, a standardized coordinate system where axes are oriented toward the patient's right, anterior, and superior directions. Scan bases were removed to retain only the gingival and dental structures. Meshes average $92,201 \pm 28,140$ vertices and

[3] `https://ditto.ing.unimore.it/bits2bites`
[4] `https://github.com/AImageLab-zip/Bits2Bites`

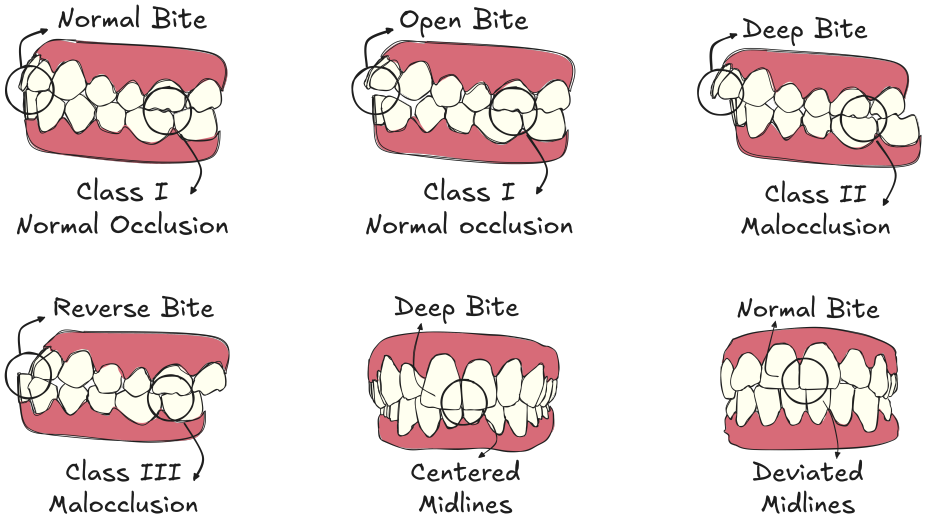

Fig. 2: Example of different occlusal classes present in our dataset.

$182,444 \pm 55,862$ faces, with bounding-box dimensions of approximately 65.9 ± 4.12 mm (width), 53.84 ± 4.5 mm (depth), and 17.9 ± 1.9 mm (height). The mean mesh surface area is $\sim 3,780 \pm 409\,\text{mm}^2$.

Scans were acquired using two different intra-oral scanners, Carestream and 3Shape TRIOS, to capture variability in acquisition technologies. The scans included were selected randomly without any filtering criteria to reflect the natural diversity and distribution observed in clinical practice.

Annotations were performed by a single orthodontic specialist with five years of experience in the field. Each scan pair includes detailed, clinically relevant occlusion labels across multiple dimensions. Sagittal classifications are provided separately for the left and right sides, following a subset of Angle's standard classification [9] (i.e., *Class I*, *Class II edge-to-edge*, *Class II full*, *Class III*). Vertical anterior–posterior relationships are labeled as *Normal*, *Deep Bite*, *Reverse Bite*, or *Open Bite*. Transverse relationships are identified as *Normal*, *Cross Bite*, or *Scissor Bite*, using reference teeth. Finally, midline alignment is annotated as *Centered* or *Deviated*. Fig. 2 provides illustrations of these different characteristics. This multi-label annotation scheme enables clinically meaningful classification across sagittal, vertical, and transverse planes. The class distribution inside the proposed dataset is reported in Fig. 4.

Ethical Approval. Approval of all ethical and experimental procedures and protocols, as well as the release of data, was granted by the Comitato Etico di Ateneo di Ferrara under Approval No. 262/2025/Oss/UniFe.

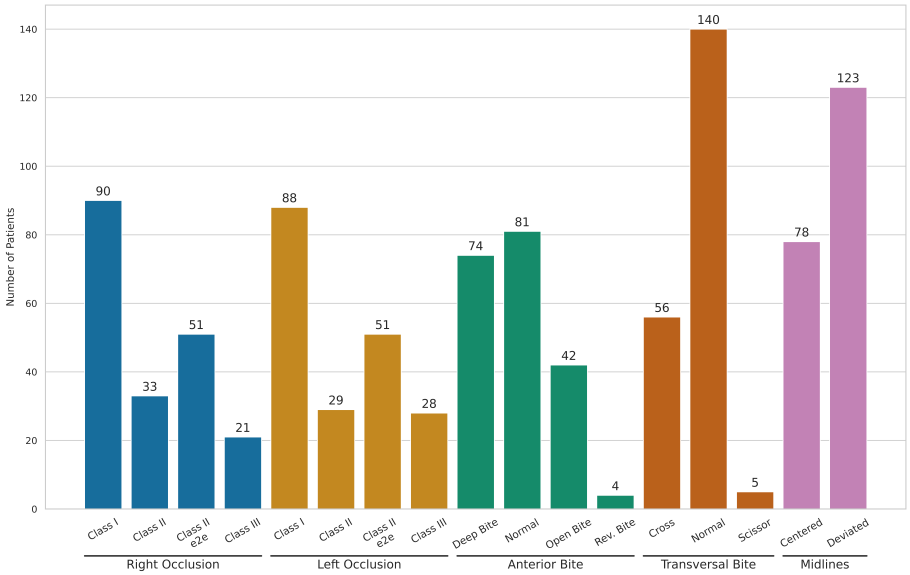

Fig. 4: Distribution of dataset classes. Distinct colors indicate different tasks.

3 Method

To address the challenge of multi-dimensional occlusion classification from intraoral scans, we developed a multi-task, point-based classification pipeline built on the open-source *Pointcept* framework [6]. Our approach jointly predicts five occlusal attributes from a single 3D point cloud representing the combined upper and lower dental arches.

Input Representation. Each sample consists of a registered pair of upper and lower intra-oral scans in STL format. Meshes are combined into a single 3D structure and converted into point clouds, where each point is represented by its *xyz* coordinates. This representation is optionally enriched with one-hot encoded per-tooth landmark features (Fig. 3) to capture anatomical context better. These landmarks are not manually annotated, but automatically predicted using the publicly available[5] state-of-the-art method from the 3DTeethLand challenge [15]. The dataset was split into five folds of 40 scans each to support a robust 5-fold cross-validation schema.

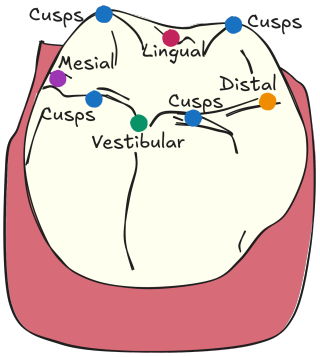

Fig. 3: Landmarks employed as additional input features.

[5] `https://github.com/nnistelrooij/3dteethland`, *final_test_phase* commit.

Preprocessing and Augmentation. To improve generalization, our training procedure leverages a carefully designed data augmentation pipeline. Each scan is normalized to a unit sphere, randomly scaled in all directions ($[0.95, 1.05]$), shifted (± 0.02 mm), rotated ($\pm 18°$ on z-axis), and subjected to random dropout (50% of points with 50% probability). Finally, it is processed with grid sampling (voxel size 0.01 mm) and converted to a tensor. Validation and testing only apply normalization and grid sampling.

Task Formulation. We frame occlusion analysis as a multi-task classification problem with a single common backbone and five independent output heads, each corresponding to: (*i*) right sagittal classification (3 classes), (*ii*) left sagittal classification (3 classes), (*iii*) anterior vertical bite type (4 classes), (*iv*) transverse bite type (3 classes), and (*v*) midline alignment (2 classes). Each head performs categorical classification using ground-truth annotations.

Model Architecture. We evaluated two state-of-the-art point cloud backbones, PointTransformerV3 [18] and SPUNet [12], both already integrated within the Pointcept framework [6]. Our benchmark formulates occlusion analysis as a multi-task classification problem, where a single shared feature extractor is followed by five independent task-specific classification heads. Each head is implemented as a two-layer multilayer perceptron (MLP) with a final softmax activation. To assess the effectiveness of this approach, Sec. 4 also conducts an ablation study comparing the multi-task learning (MTL) setup with a single-task learning (STL) strategy that trains a separate dedicated model per task.

Training Configuration. Training was performed for 200 epochs with a batch size of 8. The PointTransformerV3 models used the AdamW optimizer (learning rate 1×10^{-4}, weight decay 0.01) with a cosine annealing scheduler, while SPUNet models employed the SGD optimizer (learning rate 1×10^{-3}, weight decay 0.01) with a multistep schedule. Mixed-precision training and gradient clipping set to 1.0 were used to stabilize learning for both backbones and in all the training performed. The total loss is computed as the unweighted mean of the five task-specific losses. Each of these task-specific losses is a cross-entropy function with pre-computed class weights to address label imbalance. The exact weights are available in the source code inside the configuration files.

Evaluation Protocol. We adopted a 5-fold cross-validation scheme. In each fold, 160 scans were used for training and 40 for testing. No dedicated validation split was used. Final results are reported as the mean and standard deviation of per-task classification scores across the five folds.

4 Experiments

Experimental Setup and Metrics. We conducted our experiments following a 5-fold cross-validation protocol. For each fold, models were trained on four partitions and evaluated on the remaining held-out split, ensuring that every sample is used for testing exactly once. We evaluated two different backbones, PointTransformerV3 and SPUNet, to provide a robust benchmark for future research.

Table 1: **Ablation study on input features.** All classification metrics are macro-averaged across the five occlusal tasks and reported as mean ± std (%) over the 5 cross-validation folds. Inference time is the average time in seconds to process a single scan.

Input Features	Model	Accuracy	Precision	Recall	F1-Score	Time (s)
Mesh		0.69 ± 0.03	0.62 ± 0.02	0.61 ± 0.04	0.60 ± 0.03	0.11
Landmarks	PointTr.V3	0.70 ± 0.04	0.62 ± 0.04	0.63 ± 0.05	0.61 ± 0.04	0.04
Mesh + Landmarks		$\mathbf{0.71 \pm 0.03}$	$\mathbf{0.64 \pm 0.03}$	$\mathbf{0.64 \pm 0.02}$	$\mathbf{0.63 \pm 0.03}$	0.11
Mesh		0.64 ± 0.01	0.56 ± 0.03	0.58 ± 0.03	0.56 ± 0.04	0.05
Landmarks	SPUNet	0.60 ± 0.02	0.56 ± 0.06	0.56 ± 0.06	0.58 ± 0.05	**0.02**
Mesh + Landmarks		0.65 ± 0.01	0.59 ± 0.05	0.61 ± 0.04	0.58 ± 0.05	0.05

Both backbones were trained using identical data processing and augmentation strategies to ensure a fair and direct comparison.

Given the significant class imbalance inherent in clinical dental datasets, we selected the macro-averaged *F1-score* as our primary evaluation metric. This metric provides a balanced measure of a model's performance by calculating the F1-score for each class independently and then averaging them. In our context, such an approach ensures a more informative evaluation w.r.t. using the overall accuracy. For a more comprehensive analysis, particularly in our ablation studies, we also report accuracy, precision, recall, and model inference time.

On the Impact of Input Features. To determine the optimal input representation, we first conducted an ablation study on the input features. We compared the performance of models trained using three different input configurations: (*i*) the raw 3D mesh only, (*ii*) automatically-predicted landmark coordinates only, and (*iii*) a combination of both mesh and landmark features.

The results, summarized in Tab. 1, show that combining mesh and landmark features yields the best overall performance for both backbones, with PointTransformerV3 achieving the highest F1-score of 0.63 ± 0.03. Interestingly, using only landmark coordinates as input provides results that are only marginally lower than using the full mesh. This is a noteworthy finding, as the landmark-only models are exceptionally efficient; for instance, training takes approximately 15 minutes, compared to over 2 hours for models that process the entire mesh. Despite the efficiency of the landmark-only approach, to maximize performance, we chose configuration (*iii*) for all subsequent experiments.

Multi-Task vs. Single-Task Learning. Having established the optimal input features, we then evaluated the difference in performance between multi-task learning (MTL), i.e., a single backbone with a head for each different task, versus a single-task learning (STL) approach, i.e., a dedicated network (backbone + head) for each task. For this comparison, we trained five separate single-task models (one for each of our classification tasks) and evaluated their performance against that of our single, unified multi-task model.

As shown in Tab. 2, the STL approach, where each task is handled by a specialized model, achieves superior performance in terms of F1-score. However,

Table 2: **Ablation study on Multi-Task Learning (MTL) vs. Single-Task Learning (STL).** All classification metrics are macro-averaged across the five occlusal tasks and reported as mean ± std (%) over the 5 cross-validation folds. Inference time is the average time in seconds to process a single scan.

Model	Learning Strategy	Accuracy	Precision	Recall	F1-Score	Time (s)
PointTr.V3	Single-Task (STL)	0.72 ± 0.13	0.66 ± 0.14	0.65 ± 0.14	0.64 ± 0.13	1.10
	Multi-Task (MTL)	0.71 ± 0.03	0.64 ± 0.03	0.64 ± 0.02	0.63 ± 0.03	0.11
SPUNet	Single-Task (STL)	0.67 ± 0.14	0.61 ± 0.13	0.61 ± 0.14	0.60 ± 0.13	0.50
	Multi-Task (MTL)	0.65 ± 0.01	0.59 ± 0.05	0.61 ± 0.04	0.58 ± 0.05	0.05

Table 3: **Per-task F1-score (%) across occlusal classification tasks.** Results are macro-averaged over 5-fold cross-validation and reported as mean ± std (%).

Model	Strategy	Right Occl.	Left Occl.	Anter. Bite	Tran. Bite	Midline	Avg.
PointTr.V3	STL	0.71 ± 0.05	0.67 ± 0.07	0.77 ± 0.14	0.59 ± 0.10	0.49 ± 0.06	0.64 ± 0.13
	MTL	0.69 ± 0.05	0.68 ± 0.04	0.74 ± 0.14	0.57 ± 0.12	0.46 ± 0.05	0.63 ± 0.03
SPUNet	STL	0.60 ± 0.02	0.57 ± 0.02	0.78 ± 0.13	0.58 ± 0.14	0.48 ± 0.04	0.62 ± 0.14
	MTL	0.54 ± 0.07	0.59 ± 0.04	0.68 ± 0.15	0.61 ± 0.15	0.51 ± 0.08	0.60 ± 0.13

this gain comes at a significant cost in computational resources and complexity. The STL strategy requires training and maintaining five distinct models per backbone, resulting in an increase in total training time and inference overhead compared to the unified MTL model. Tab. 3 provides a more granular, per-task breakdown of the F1-scores, confirming the strong performance of the STL models across the individual tasks.

These results present a clear trade-off: the MTL framework offers an efficient and scalable solution well-suited for clinical application where speed may be critical, while the STL approach can provide higher accuracy if computational cost is not a primary constraint. Across all experiments, the PointTransformerV3 backbone consistently outperformed SPUNet, establishing it as the more robust architecture for this problem domain.

Qualitative and Error Analysis. To better illustrate model performance beyond quantitative metrics, Fig. 5 shows example predictions from our test set, highlighting both successful classifications and common failure modes. These visualizations provide insight into the models' ability to interpret complex inter-arch relationships and offer a qualitative understanding of their predictive behavior in challenging clinical cases.

5 Conclusion

In this paper, we introduced *Bits2Bites*, a novel benchmark for occlusal classification from intra-oral scans. We provided the first public dataset of 200 paired

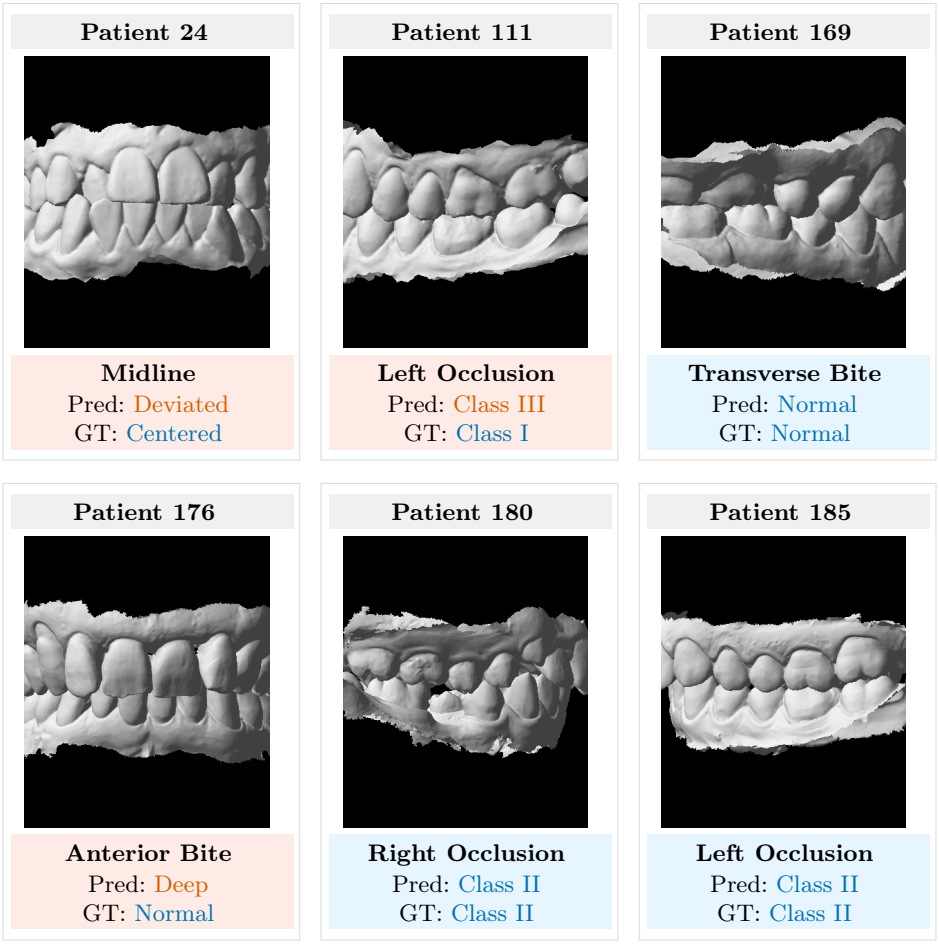

Fig. 5: Qualitative analysis of PointTransformerV3 model predictions on various occlusal classification tasks. The figure showcases both correct classifications , where the model's prediction matches the ground truth, and failures , where the model misclassifies one task in the scan. Each example compares the model's prediction (Pred) with the expert-annotated ground truth (GT) for a specific patient from the test set.

IOS scans with detailed, multi-dimensional clinical annotations. Our evaluation of state-of-the-art point cloud backbones within a multi-task learning framework demonstrates the feasibility of directly predicting multiple occlusal attributes from raw 3D point clouds. The results of our experiments lay the groundwork for developing automated tools that can assist orthodontists in diagnosis and treatment planning.

Future work will proceed in three main directions. First, we plan to expand the dataset to include a larger and more diverse cohort of patients, capturing

a wider range of rare malocclusions. Second, we will validate the clinical annotations by involving multiple experts to establish inter-rater reliability, further strengthening the quality of the ground truth. Finally, once the dataset is enriched, support for the previously merged Class II edge-to-edge and Class II full sagittal classifications, as well as for tooth-level identification in crossbite and scissor bite cases, will be reinstated and fully integrated.

References

1. Ben-Hamadou, A., Smaoui, O., Rekik, A., Pujades, S., Boyer, E., Lim, H., Kim, M., Lee, M., Chung, M., Shin, Y.G., et al.: 3DTeethSeg'22: 3D Teeth Scan Segmentation and Labeling Challenge. arXiv preprint arXiv:2305.18277 (2023)
2. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., Van Nistelrooij, N., Van Lierop, P., Xi, T., Liu, Y., et al.: Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. IEEE Transactions on Medical Imaging (2024)
3. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volumes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
4. Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., Minafra, P., Anesi, A., Grana, C.: Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. IEEE Access (2022)
5. Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., Grana, C.: Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
6. Contributors, P.: Pointcept: A Codebase for Point Cloud Perception Research. `https://github.com/Pointcept/Pointcept` (2023)
7. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al.: A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nature Communications **13**(1) (2022)
8. Di Bartolomeo, M., Pellacani, A., Bolelli, F., Cipriano, M., Lumetti, L., Negrello, S., Allegretti, S., Minafra, P., Pollastri, F., Nocini, R., et al.: Inferior Alveolar Canal Automatic Detection with Deep Learning CNNs on CBCTs: Development of a Novel Model and Release of Open-Source Dataset and Algorithm. Applied Sciences **13**(5) (2023)
9. Graber, L.W., Vanarsdall, R.L., Vig, K.W., Huang, G.J.: Orthodontics: Current Principles and Techniques (1994)
10. Isensee, F., Kirchhoff, Y., Kraemer, L., Rokuss, M., Ulrich, C., Maier-Hein, K.H.: Scaling nnU-Net for CBCT Segmentation. In: Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data (2025)
11. Juneja, M., Saini, S.K., Kaur, H., Jindal, P.: Application of Convolutional Neural Networks for Dentistry Occlusion Classification. Wireless Personal Communications **136**(3) (2024)
12. Liu, X., Liu, X., Liu, Y.S., Han, Z.: SPU-Net: Self-Supervised Point Cloud Upsampling by Coarse-to-Fine Reconstruction with Self-Projection Optimization. IEEE Transactions on Image Processing **31** (2022)
13. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access (2024)

14. Ma, Q., Sun, G., Tombak, G.I., Jain, S., Huber, N.B., Gool, L.V., Konukoglu, E.: Video Foundation Model for Medical 3D Segmentation. In: Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data (2025)
15. van Nistelrooij, N., Vinayahalingam, S.: ToothInstanceNet: Comprehensive Information from Intra-oral Scans by Integration of Large-Context and High-Resolution Predictions. In: Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data (2025)
16. Rekik, A., Ben-Hamadou, A., Smaoui, O., Bouzguenda, F., Pujades, S., Boyer, E.: TSegLab: Multi-stage 3D dental scan segmentation and labeling. Computers in Biology and Medicine **185** (2025)
17. Wang, C., Zhang, Y., Wu, C., Liu, J., Huang, X., Wu, L., Wang, Y., Feng, X., Lu, Y., Wang, Y.: MMDental-A multimodal dataset of tooth CBCT images with expert medical records. Scientific Data **12**(1) (2025)
18. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point Transformer V3: Simpler, Faster, Stronger. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
19. Zou, B., Wang, S., Liu, H., Sun, G., Wang, Y., Zuo, F., Quan, C., Zhao, Y.: TeethSEG: An Efficient Instance Segmentation Framework for Orthodontic Treatment based on Anthropic Prior Knowledge. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)