# Multi-Class Dental CBCT Segmentation in Data-Constrained Scenarios Through Transformers

Rafael C. Giménez-Aguilar[1*], Sergio Paraíso-Medina[1], Miguel García-Remesal[2], Guillermo Jesús Pradíes-Ramiro[3], Monica Bonfanti-Gris[3], Raul Alonso-Calvo[1*]

[1] Universidad Politécnica de Madrid, Biomedical Informatics Group, Department of Computer Languages and Systems and Software Engineering, Madrid (Spain)
[2] Universidad Politécnica de Madrid, Biomedical Informatics Group, Department of Artificial Intelligence, Madrid (Spain)
[3] Complutense University of Madrid, Department of Bucofacial Prostheses, Madrid (Spain)

* Corresponding author: rafael.gimenez@upm.es (R. C. Giménez-Aguilar), raul.alonso@upm.es (R. Alonso-Calvo).

Received 11 September 2023 | Accepted 7 February 2025 | Published 12 March 2025

## Abstract

Accurate segmentation of dental structures from cone-beam computed tomography (CBCT) images has become an active research field due to the widespread use of this technology in clinical practice. In recent years, contributions have shifted from traditional computer vision methods to deep learning-based approaches. However, most of these works are based solely on convolutional neural networks (CNNs), whereas the image segmentation state-of-the-art is currently moving towards attention-based architectures. Furthermore, contributions on dental CBCTs predominantly present methods focused on a single object category, mainly teeth. In this article we tackle the segmentation of multiple oral structures by implementing previously unutilized query-based segmentation transformers. The proposed method achieves similar results to the state-of-the-art, especially on tooth segmentation, while employing a considerably smaller training dataset than prior contributions.

## Keywords

Dental CBCT, Deep Learning, Instance Segmentation, Multi-class Segmentation, Transformer.

DOI: 10.9781/ijimai.2025.03.003

## I. Introduction

Cone-Bean computed tomography (CBCT) has become a widespread diagnostic imaging modality in dental practice, with some surveys reporting up to 76% of clinics having one machine, and 33% performing at least one scan every year [1].

Due to the ability of CBCT to yield a volumetric reconstruction of the patient's mouth, it is commonly employed in implant planning, impacted tooth and temporomandibular joint evaluation [2]. However, the characteristics of the machinery and the energy spectrum used in this modality lead to a higher expression of image artifacts. This is especially noticeable in cases where the patient has some form of dental implant constructed with highly attenuating materials. As a result of these phenomena, the manual creation of volumetric masks of dental structures such as teeth, bone, or sinuses, is often time-consuming and requires an experienced dental practitioner.

Automatic segmentation of tomographic modalities such as computed tomography (CT) and magnetic resonance imaging (MRI), have been extensively researched, employing both classical computer vision and machine learning approaches [3]. Yet, most of these methods cannot be directly translated to CBCT due to the prevalence of the previously mentioned artifacts. In addition, they have smaller spatial resolution than clinical CT and MRI machines.

Segmentation, like most computer vision tasks, has recently seen a shift from traditional algorithms to deep learning-based methods [4]. This was further accentuated by the advent of convolutional neural networks (CNNs), which replaced or were combined with previous methods in tasks such as image classification, object detection or image segmentation [5]. Such transition can also be observed in the CBCT segmentation literature, where contributions have gradually moved from traditional computer vision algorithms [6]–[13] to machine learning methods [14]–[31], especially CNNs [17]–[31].

Despite that, the computer vision field is now seeing another paradigm shift with the introduction of Transformers [32]. These neural networks, initially conceived for natural language processing (NLP) tasks, have recently been adapted to be used with images [33] and have shown comparable, and in some cases superior, performance compared to CNNs. Furthermore, as a result of recent research, unified architectures have been described that perform all three types of

Please cite this article as: R. C. Giménez-Aguilar, S. Paraíso-Medina, M. García-Remesal, G. J. Pradíes-Ramiro, M. Bonfanti-Gris, R. Alonso-Calvo. Multi-Class Dental CBCT Segmentation in Data-Constrained Scenarios Through Transformers, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 6, pp. 52-60, 2026, http://doi.org/10.9781/ijimai.2025.03.003

segmentation tasks simultaneously (semantic, instance, and panoptic) in what has been referred to as universal segmentation [34].

The absence of contributions leveraging these new advances in dental CBCT segmentation motivated this work. Thus, in this paper we present a multi-class method that performs instance segmentation of teeth and metal implants. In addition, we perform semantic segmentation of maxillary and mandible bone, and maxillary sinuses. The main contributions of this work are:

(i) We propose the use of query-based instance segmentation Transformers which show state-of-the-art performance in common segmentation benchmarks (COCO [35], Cityscapes [36], ADE20K [37] and Mapillary Vistas [38]), on this problem.

(ii) We produce a multi-class model, tackling four types of oral structures, whereas most of the literature deals with one or two classes.

(iii) We trained our model on a single patient's dataset and validated it on eleven different patients, acquired using two distinct X-ray systems. Despite being trained on a smaller dataset, our model outperforms previous works trained on larger datasets and achieves performance close to the state of the art.

Our approach resorts to two components: an encoder-decoder vision Transformer, and a post-processing module that refines the segmentation outputs, providing instance segmentation in certain categories and semantic segmentation in others.

In order to further enhance the generalization capabilities of the architecture, some of its components were pre-trained on large-scale datasets such as COCO [35].

## II. Background and Related Work

This section provides a survey of previous contributions regarding the automatic segmentation of oral structures. It also provides an overview of Vision Transformers, primarily focused on segmentation tasks.

### A. CBCT Automatic Oral Structure Characterization

Segmentation of oral structures such as teeth, implants or bone has been widely studied, with most contributions performing semantic segmentation, separating a whole oral structure – such as the denture – from the rest; or instance segmentation, which involves detecting and segmenting individual entities of certain category, as in the case of teeth.

This literature could be taxonomically organized into contributions employing traditional computer vision (CV) approaches, and those leveraging datasets to train machine learning models.

#### 1. Traditional CV Approaches

Among these contributions, the most prevalent techniques found in the literature are level set methods [8], [9], [10], [13], statistical shape models (SSMs) [6], [7], [9] and watershed segmentation [12].

Level set methods, a form of active contour methods, extract borders by iteratively modifying the zero-cut of a high-dimensional function. Gao and Chae [8] use this technique leveraging prior information from contiguous slices for individual tooth segmentation. This approach was further developed by Yau et al. [10], Ji et al. [13] and Gan et al. [11].

Alternatively, SSMs follow the notion of deforming a high-dimensional structure which is often a template constructed from a dataset. This was first applied to mandible segmentation by Lamecker et al. [6], and further improved by segmenting the mandibular nerve in Kainmueller et al. [7]. Duy et al. [9] continued this line of research by individual tooth segmentation.

Watershed, a classic non-semantic segmentation technique, was recently used by Fan et al. [12] to dilate manual marked regions to perform mandible segmentation.

#### 2. Machine Learning Approaches

Machine learning is a branch of artificial intelligence based on methods aimed at emulating human learning processes, gradually improving their performance by learning from data. Therefore, their performance is tightly coupled with the amount and quality of training datasets. The most used approaches for CBCT segmentation found in the literature were Markov random fields (MRFs), mean-shift algorithms, random forests and CNNs.

MRFs are probabilistic graphical models, designed to derive information from pixels to cluster or segment them. They have been optimized through graph cutting algorithms for tooth segmentation by Hiew et al. [14], and later combined with SSMs by Keustermans et al. [15].

Mortaheb et al. [16] employed an unsupervised non-parametric clustering algorithm for teeth segmentation, using the CIELUV color space as their feature space.

Alternatively, Wang et al. [39] implemented a sequential combination of random forest metaclassifiers, to segment mandibles within CBCT images.

Nonetheless, the most prevalent approach in CV and, particularly, in CBCT segmentation research is CNN based methods. Most of these publications either repurpose the use an established CNN architecture or propose some variation on their topology, with the most common being U-Net [40], MS-D [41] and Mask R-CNN [42]. Most contributions perform a single information extraction task, such as: denture segmentation [19], [20], [28], [29], [31], tooth instance segmentation [17], [21], [22], metal segmentation [18] and sinus segmentation [25]. Only five works were found to address two or more categories [23], [24], [26], [27], [30]. A summary of the contributions using CNNs can be found in Table I.

TABLE I. Overview of Contributions Employing CNNs

| Contribution | Approach | Dataset size |
|---|---|---|
| Cui et al. [17] | Edge encoder-decoder + 3D Mask R-CNN | 20 patients |
| Hegazy et al. [18] | U-Net | 5 patients |
| Lee et al. [19] | UDS-Net (U-Net + novel dense blocks) | 102 patients |
| Rao et al. [20] | U-Net + Dense Conditional Random Field | 86 images |
| Chen et al. [21] | V-Net + watershed | 25 patients |
| Jang et al. [22] | One-step detection + U-Net segmentation | 97 volumes |
| Wang et al. [23] | MS-D | 30 patients |
| Zheng et al. [24] | Dense U-Net | 20 patients |
| Morgan et al. [25] | 3D U-Net | 83 volumes |
| Cui et al. [26] | Encoder-decoder CNN ensemble | 4215 patients |
| Dot et al. [27] | U-Net | 603 patients |
| Hu et al. [28] | DSFNet | 150 patients |
| Jing et al. [29] | USCT | 85 patients |
| Nogueira-Reis et al. [30] | U-Net ensemble | 30 patients |
| Wang et al. [31] | Trans V-Net (V-Net with cross-attention skip connections) | 150 patients |

### B. Vision Transformers

Attention mechanisms initially emerged to overcome the loss of context experienced by encoder-decoder Recurrent Neural Networks (RNNs) in sequence-to-sequence (seq2seq) tasks [43]. In that setting, it served as a way of combining the hidden states of the encoder weighing each contribution in terms of the prior hidden state in the decoder. It wasn't until the introduction of the Transformer by Vaswani et al. [44] that the attention mechanism was employed as the main building block of a neural network.

Attention itself can be understood in terms of classical regression [45], [46], where a set of outputs is obtained by comparing the similarity between some set queries and key-value pairs. In Transformers, however, these queries, keys and values are projections of the inputs given to the layer. In general, they can be represented as

$$Q = X_q W_q \in R^{n \times d_k} \tag{1}$$

$$K = X_{k,v} W_k \in R^{m \times d_k} \tag{2}$$

$$V = X_{k,v} W_v \in R^{m \times d_v} \tag{3}$$

where Q, K and V are the matrices corresponding to the queries, keys and values respectively; $W_q \in R^{d \times d_k}$, $W_k \in R^{d' \times d_k}$ and $W_v \in R^{d' \times d_v}$ are learnable weights that produce the projections of the inputs; $X_q \in R^{n \times d}$ is the set of $n$ vector embeddings that form the queries and $X_{k,v} \in R^{m \times d'}$ are $m$ vector embeddings used for the keys and values.

When the same input is used in queries, keys and values, this mechanism is termed self-attention, otherwise it may be referred to as cross-attention. Nonetheless, these attention blocks are often implemented as multi-head attention, where $h$ attention mechanisms are performed in parallel, and their resulting embeddings are concatenated along the embedding dimension and projected with another trained matrix $W_0 \in R^{hd_v \times d'_v}$.

Another relevant component of Transformers' attention mechanism is positional encoding. To preserve information about the order of the sequence and the tokens in the embeddings, an encoding is added or concatenated to the input embedding before performing the attention operation. This embedding can be found as a fixed value derived from sinusoids [44] or as a set of learnable parameters [47].

As a final consideration, it is important to mask the attention of a Transformer decoder to avoid prior queries attending to keys belonging to the latter ones. This is often done by adding a matrix such that $m_{i,j} = -\infty$ for any $i > j$. These contributions will be zeroed after performing the softmax.

A Transformer block is usually created by combining self or cross multi-head attention operations with residual connections, multi-layer perceptrons and normalization operators, concretely, layer normalization [48].

The unprecedented scalability of Transformers and their zero-shot performance when pre-trained on massive datasets in NLP, led to the creation of the Vision Transformer (ViT) [33], an initial approach to adapting the standard Transformer to images. ViT and other basic implementations of the Transformer in images lacked hierarchical feature representation. This is a desirable trait for backbone networks or feature extractors in several popular architectures. This was first tackled by the Pyramid Vision Transformer (PVT) [49], with later modifications to improve the scalability of these approaches with image size in the Swin Transformer [50]. Based on this, Transformers could be used as a drop-in replacement for CNN feature extractors and employed to increase the performance of previously fully convolutional architectures in multiple computer vision tasks, such as classification, detection, or segmentation.

Regarding the segmentation task, Liu et al. [51] propose a comprehensive taxonomy to classify recent Transformers. According to it, architectures can be grouped into patch-based or query-driven ones. The prior naïvely implemented the initial concept of the Vision Transformers using patch embeddings to perform per-pixel classification. Conversely, query-based models employ object or mask queries that are processed along with image features to produce output embeddings that determine the bounding box, class, and mask of the detections.

## III. Materials and Methods

In this section we discuss the methods we used for training the classification models. We also describe the experiments we conducted to assess the contributions we report on this paper. First, we provide an overview of the dataset used in this study. Then, we describe the adopted deep learning architecture, including its training and the experiments we conducted to assess its performance.

### A. Data Acquisition

The dataset used in this project encompasses 12 patients, whose images were acquired with two different systems. Data from a single patient was used for training, while the rest of the data was used at the validation stage. Two patients were scanned using a Carestream CS 8100 3D; the training volume was obtained with 90 kVp, 30 mAs and a spatial resolution of 0.15 mm, and the other validation images with 85 kVp, 70.9 mAs and a resolution of 0.16 mm. The other patients within the validation images were captured on a Vatech PHT-35LHS with 94 kVp, 114.8 mAs and a resolution of 0.2 mm.

All image annotations were manually performed by an expert orthodontist using COCO Annotator [52]. The target categories were tooth, metal, bone (mandible and maxilla), and maxillary sinus. Teeth and metal implants were annotated as instances, while bone and maxillary sinuses were represented with a single semantic mask in the image. The training dataset comprised of a whole patient volume, made from 384 annotated slices of 512-pixel width and height, with 4043 annotations. The validation dataset contained 25 slices from 11 patients with a square size of 550 pixels and 316 annotations.

### B. Deep Learning Implementation

Our network design is based on the MaskFormer [53] family of architectures, a type of query-based segmentation model. Their meta structure is constructed using three modules: a feature extractor in the form of a Feature Pyramid Network (FPN) [54], a Transformer module and a segmentation module. Our implementation maintains such organization, modifying the perceptron used for classification and introducing several detection refinement steps. This is to transition from all instance detections to our desired instance and semantic outputs. Images of size H×W with C channels are input into a feature extractor in the form of an FPN made up of a Swin-S backbone and a CNN decoder. The multi-scale feature maps are used as inputs for the transformer decoder and mask logits generation. The transformer decoder makes as many detections as queries are set in the model configuration. The resulting embeddings are then used to obtain class predictions and masks in conjunction with the feature maps. Lastly all object detections are refined by a post-processing block. Fig. 1 shows an overview of this architecture.

The FPN module was built using the small variant of Swin [50] (Swin-S) as the feature extractor backbone. Such backbone produces 4 multi-resolution feature maps that are then fed into a convolutional network as described in the original FPN paper [54]. The three lower spatial dimension outputs of the FPN were used as inputs for the Transformer decoder. The last highest-dimension feature map was employed in the segmentation module.

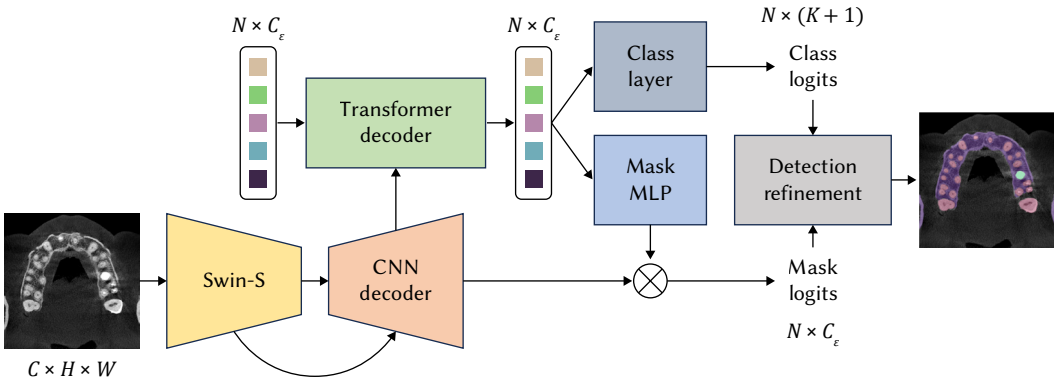



Fig. 1. Diagram of the proposed MaskFormer-based architecture.

The Transformer decoder used Mask2Former's [34] implementation, employing masked attention instead of cross-attention on the query embeddings and image features. This attention mechanism applies the previous layer's mask predictions on the current layer's attention, computed as

$$X_l = \text{softmax}(M_{l-1} + Q_l K_l^T) V_l + X_{l-1} \quad (4)$$

where $M_{l-1} \in \mathbb{R}^{N \times H_l W_l}$ is the binarized output of the previous layer resized to the resolution of the input image features ($K_l$, $V_l \in \mathbb{R}^{N \times H_l W_l}$), and $Q_l \in \mathbb{R}^{N \times C_\varepsilon}$ are the input query embeddings.

An initial set of query embeddings is iteratively refined through groups of three Transformer layers. With each group layers receiving feature maps from the FPN at 1/32, 1/16 and 1/8 of the input resolution respectively. As in its original implementation, this process is repeated $L$ times, producing a decoder with $3L$ layers. Our implementation used 100 queries ($N$) with an embedding size ($C_\varepsilon$) of 256 and 3 groups ($L$) of Transformer layers, leading to a total of 9 layers in the decoder. An overview of a single group of layers within the decoder can be seen in Fig. 2.

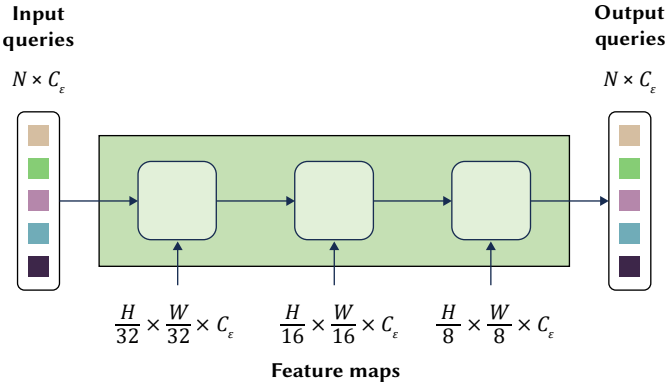



Fig. 2. Transformer layer structure repeated $L$ times to construct the transformer decoder.

The segmentation module obtains the class and mask logits through fully connected neural networks, using the Transformer decoder embedding as inputs. The multi-task loss used during training was expressed as

$$L = L_{cls} + L_{CE} + L_{Dice} \quad (5)$$

here, $L_{cls}$ represents the cross-entropy loss of the class predictor, $L_{CE}$ is the binary cross-entropy loss of the masks and $L_{Dice}$ is the Dice loss [55] of the masks.

The resulting logits are refined to produce output predictions through a two-stage process. In the first stage, the class and mask scores are obtained through the softmax and sigmoid functions respectively. From the initial 100 queries, predictions are pruned if they belong to the background (null) class, or if their score is under a certain threshold. In the second stage, the instance segmentation predictions are subjected to two operators: a connected-component labeling algorithm based on their class, and non-maximum suppression of the resulting blobs.

During the connected-component labeling part of the post-processing, if the predictions belong to either bone or maxillary sinus, their masks are separated into individual blobs using the Spaghetti [56] algorithm. All the resulting binary masks are then subjected to non-maximum suppression using their intersection-over-union as the deciding metric, with a threshold of 0.3. Lastly, masks belonging to bone and sinus are merged back into a single prediction to accomplish semantic segmentation in these categories and instance segmentation in teeth and metal.

### C. Model Training Settings

Aside from the linear classifier used for class prediction, all parameters in the model were initialized from a baseline pre-trained with the COCO dataset [35]. During training, the Swin Transformer was used as a fixed feature extractor, only modifying the pixel decoder, Transformer decoder and fully connected layers in the segmentation module.

To increase the generalization ability of the model, a series of geometric and intensity transformations were applied during the training phase to resemble differences found amongst patients and tomographic systems. Each image was subjected to random horizontal flipping, affine transforms (translation, scaling, and rotation), median blur and gamma correction.

The model was trained for 75 epochs with a batch size of 16, using the AdamW [57] optimizer, a learning rate of $1 \times 10^{-4}$ and an L2 weight decay of $5 \times 10^{-2}$. All learning rates were reduced by a factor of 0.1 at 90% and 95% of the training.

## IV. Results

The evaluation was done on the validation dataset described in section III.A of the Materials and Methods. We assessed the performance of our model on two sets of metrics: object detection and segmentation. Regarding object detection, all average precision COCO [35] metrics were used. Segmentation evaluation was performed through the Dice similarity coefficient or DSC (equivalent to the F1 score), precision and recall as overlap metrics, and the Hausdorff distance (HD) and average symmetric surface distance (ASSD) as boundary metrics. Both sets of metrics will be presented in terms of all detections and on a per-label basis. Additionally, a set of ablation studies will be presented in the following section to assess the impact of the data augmentation scheme employed.

TABLE II. Multi-Class Segmentation Results

| Category | DSC (%) | Precision (%) | Recall (%) | HD (mm) | ASSD (mm) |
|---|---|---|---|---|---|
| Tooth | 91.0 ± 6.3 | 97.8 ± 3.1 | 85.9 ± 10.1 | 0.86 ± 0.99 | 0.25 ± 0.16 |
| Metal | 89.3 ± 4.2 | 98.1 ± 4.1 | 82.5 ± 7.9 | 0.48 ± 0.15 | 0.19 ± 0.07 |
| Bone | 87.9 ± 6.3 | 83.3 ± 8.5 | 93.6 ± 6.3 | 14.51 ± 13.67 | 1.65 ± 3.46 |
| Sinus | 91.1 ± 1.8 | 91.0 ± 10.6 | 92.1 ± 7.2 | 3.03 ± 1.09 | 0.57 ± 0.01 |
| | 90.7 ± 6.2 | 96.6 ± 5.6 | 86.3 ± 9.9 | 1.97 ± 5.45 | 0.36 ± 1.06 |

Average precision metrics used the intersection-over-union (IoU) of binary masks as thresholding values. The mean average precision across categories (AP) resulted in 0.61, with the AP at IoU thresholds of 0.5 and 0.75 ($AP_{50}$ and $AP_{75}$) yielding 0.89 and 0.72 respectively. AP results for each class were 0.69 for teeth, 0.39 for metal, 0.61 for bone and 0.75 for maxillary sinuses.

Segmentation metrics pertaining to boundaries were computed using each acquisition's spatial resolution, thus being expressed in millimeters (mm). All previously presented metrics are given per structure category in Table II, where each value represents the average over all detections within the validation set.

### A. Ablation Study

This section presents detection and segmentation metrics of models trained with different data augmentation schemes to gauge their influence. All measurements use the model with no train-time augmentation as the baseline and compare it to using solely geometric transformations (horizontal flipping and affine transformations) and/or intensity transformations (median blur and gamma correction).

A comparison between detection metrics amongst the data augmentation transforms is given in Table III, where each value represents the average result of all detection classes. Table IV provides a comparison between said data augmentation transforms, by presenting the mean average precision in a per-class basis.

TABLE III. Detection Ablation Results

| | Transforms | | | |
|---|---|---|---|---|
| | None | Geometric | Intensity | All |
| **AP** | 0.46 | 0.51 | 0.56 | 0.61 |
| **$AP_{50}$** | 0.80 | 0.85 | 0.84 | 0.89 |
| **$AP_{75}$** | 0.47 | 0.58 | 0.68 | 0.72 |

TABLE IV. Per-Class Mean Average Precision Results

| Category | Transforms | | | |
|---|---|---|---|---|
| | None | Geometric | Intensity | All |
| Tooth | 0.65 | 0.69 | 0.64 | 0.69 |
| Metal | 0.21 | 0.31 | 0.31 | 0.39 |
| Bone | 0.44 | 0.55 | 0.55 | 0.61 |
| Sinus | 0.55 | 0.50 | 0.75 | 0.75 |

Table V presents the segmentation results on the validation dataset averaged over all classes when training the model using differing data augmentation transforms.

TABLE V. Segmentation Ablation Results

| | Transforms | | | |
|---|---|---|---|---|
| | None | Geometric | Intensity | All |
| **DSC (%)** | 90.7 ± 6.8 | 91.6 ± 5.8 | 90.3 ± 6.5 | 90.7 ± 6.2 |
| **Precision (%)** | 93.0 ± 9.7 | 94.6 ± 7.6 | 94.6 ± 9.3 | 96.6 ± 5.6 |
| **Recall (%)** | 89.7 ± 8.2 | 89.6 ± 8.3 | 87.6 ± 9.1 | 86.3 ± 9.9 |
| **HD (mm)** | 2.11 ± 4.46 | 2.28 ± 5.88 | 2.40 ± 6.53 | 1.97 ± 5.45 |
| **ASSD (mm)** | 0.35 ± 0.55 | 0.36 ± 0.87 | 0.38 ± 0.89 | 0.36 ± 1.06 |

## V. Discussion

Automatic segmentation of oral structures in CBCTs is becoming increasingly relevant in digital dentistry. Deep learning approaches, especially CNNs, have shown strong performance when segmenting objects such as teeth [17], [19], [20], [21], [22], [26]. However, current trends image segmentation seems to be moving towards other deep learning paradigms, primarily attention-based networks, or Transformers [58].

Most recent contributions to 2D and 3D segmentation of CBCTs explore the creation of deep learning-based methods that partially or entirely use CNNs and employ a significant amount of data for training. Our study explores Transformer-based alternatives and assesses their generalization ability when trained with an extremely limited number of patients. Furthermore, as many authors focus on a single segmentation task and category, we strived to devise an approach that could segment multiple oral structures using different segmentation strategies.

To that end, we expanded a query-based segmentation architecture pre-trained on a large-scale image dataset, fine-tuned it for our task and post-processed its output to perform instance segmentation in two of our categories (tooth and metal) and semantic segmentation in the rest (bone and maxillary sinus).

The evaluation of our method in terms of object detection yielded promising results, accomplishing an average precision of 0.61 across classes. The lowest performance was found for metal objects. This is consistent with our expectation, as metal implant annotations share numerous spatial and morphological characteristics with teeth. However, they only account for 18.95 % of the annotations in the training dataset, compared to 69.65 % for teeth.

We surveyed recent contributions addressing tooth segmentation using deep learning techniques [17], [19], [20], [21], [22], [23], [26] to evaluate the performance of our proposed method for tooth instance segmentation. These contributions result in a mean DSC of 93.3 ± 1.4 % and an ASSD of 0.26 ± 0.09. Our model segmentations achieve only a 2.3 % reduction in DSC and 1 mm lower ASSD, albeit with higher standard deviation. Nonetheless, the boundary error in such small categories is mainly due to the low spatial resolution of dental CBCT images.

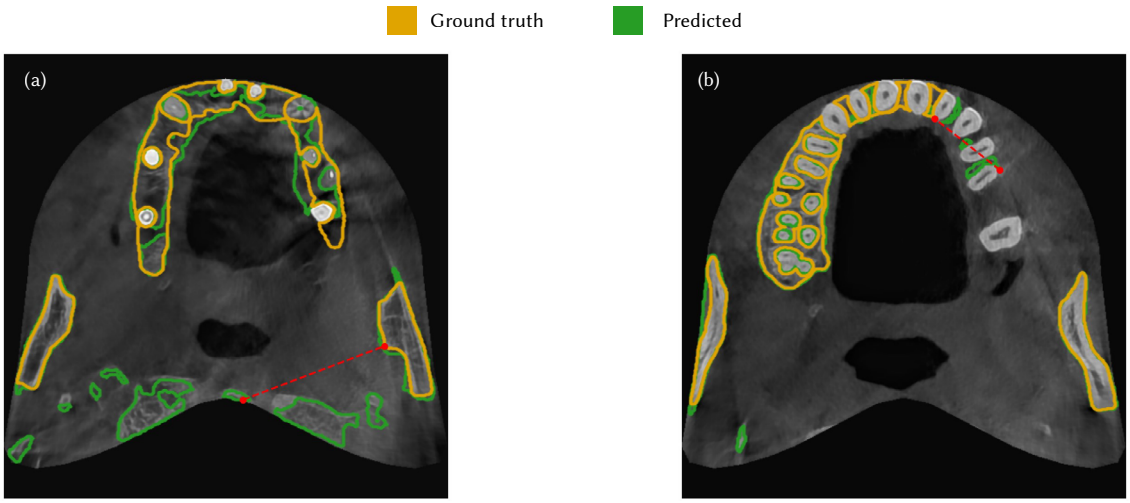



Fig. 3. Boundary errors displayed as red dotted lines, in semantic segmentation of maxillary and mandible bone. (a) Hausdorff distance resulting from segmentation of vertebrae as bone of interest. (b) Hausdorff distance arising from false positives in the predicted mask.

Semantic segmentation classes (bone and sinuses), although less numerous in terms of annotations, presented overlap scores close to those of instance segmentation, i.e., teeth and metal. However, they both show significant boundary errors, appreciable through their Hausdorff distances. When visually inspected, the origin of these flaws has been attributed to two sources: false positive regions and the segmentation of bones that do not belong to the maxilla or mandible, for instance, vertebrae. Examples of such errors can be found in Fig. 3.

Regarding the multi-class capabilities of our approach, we compare our segmentation results to that of the only three works found tackling more than a single category [23], [24], [26]. Zheng et al. [24] employ a modified and anatomically constrained version of U-Net [40], and report semantic segmentation of four object categories: lesion, material, bone, and teeth. Wang et al. [23] also address a semantic segmentation problem, using MS-D [41] to segment teeth and bone. Lastly, Cui et al. [26] uses an ensemble of encoder-decoder CNNs each addressing certain task, to obtain instance segmentation masks of teeth and bone, differentiating between maxilla and mandible in the latter. A numeric comparison of the segmentation results on common classes with the previously mentioned works is presented in Table VI. As observed in said comparison, our method outperforms prior semantic segmentation contributions in the tooth category, while additionally providing the detection of individual instances. This is especially relevant due to the nature of our training dataset. The use of a single patient volume not only restricts the morphologies the network is presented with, but it also yields a significantly reduced amount of distinct data due to the high correlation between contiguous slices within the volume.

TABLE VI. Comparison With Other Multi-Class Approaches

| Contribution | Tooth DSC (%) | Bone DSC (%) | Patients used for training |
|---|---|---|---|
| Zheng et al. [24] | 81.0 ± 6.0 | 89.0 ± 6.0 | 15 |
| Wang et al. [23] | **94.8 ± 2.0** | 93.4 ± 2.0 | 21 |
| Cui et al. [26] | 94.1 ± 1.1 | **94.5 ± 0.4** | 3172 |
| **Ours** | 91.0 ± 6.3 | 87.9 ± 6.3 | **1** |

We suggest that the accuracy discrepancies between categories in our approach may be caused by two distinct sources: the nature of the architecture and the differences in the training and validation acquisition. Regarding the first source, as the architecture is query based, it detects and segments individual instances of objects, allowing for increased performance when presented with more annotations of a certain category. In our case the most numerous ones were teeth and metal, while the other two (bone and sinuses) were less represented in the dataset. On the other hand, as the span of the scanned regions differ between the machines used for acquiring the training dataset and that of most validation patients, the model has not been allowed to learn of the existence of bone not belonging to our categories of interest, i.e., vertebrae.

The results of the ablation study on data augmentation techniques confirmed that the train-time transformations greatly improved the model's performance, especially in a low data resource scenario such as this project. The most significant enhancements being shown are in the increase of all object detection metrics and in the reduction of segmentation boundary errors.

As stated in previous contributions [26], the improvements provided by data augmentation are limited in terms of model generalization compared to the used of larger and diverse datasets. However, by leveraging the feature representation of state-of-the-art neural networks and employing the latest architectural paradigms, our method achieves promising results even with one patient in our training dataset. This is further supported, as the model can generalize not only for the morphologies of 11 other patients but also for acquisitions employing completely different machines. Examples of multi-class segmentation maps produced by the model on the validation set can be found in Fig. 4 in the Appendix.

## VI. Conclusion

This work presents a multi category approach to dental CBCT segmentation exploiting Transformer architectures, in contrast to the predominant single category – mostly teeth – CNN-based methods found in the literature. Additionally, it is achieved using an extremely limited training dataset consisting of a single patient volume.

The resulting method exploits the generalization capability of attention-based segmentation architectures pre-trained on large image datasets, fine-tunes them and processes its results to perform instance segmentation of teeth and metal implants, and semantic segmentation of maxillary bone, mandibles, and maxillary sinuses.

Our tooth segmentation results show comparable performance to those of SoTA methods even when employing much smaller training datasets. Conversely, segmentation results for less common classes such as bone and sinuses suggest that the constrained amount of data severely hinders the model's ability for generalization.

We expect future work on CBCT segmentation to benefit from moving towards a universal segmentation approach, primarily based on networks operating through attention mechanisms. Furthermore, encoder-decoder architectures such as the one employed in this contribution, could extend its capabilities with additional parallel decoders performing other tasks, such as single tooth classification. With enough data, these approaches may also be expanded to volumetric ones and could benefit from combining two types of volume representation of the same region of the patient, such as CBCT and panoramic X-rays.

## Appendix

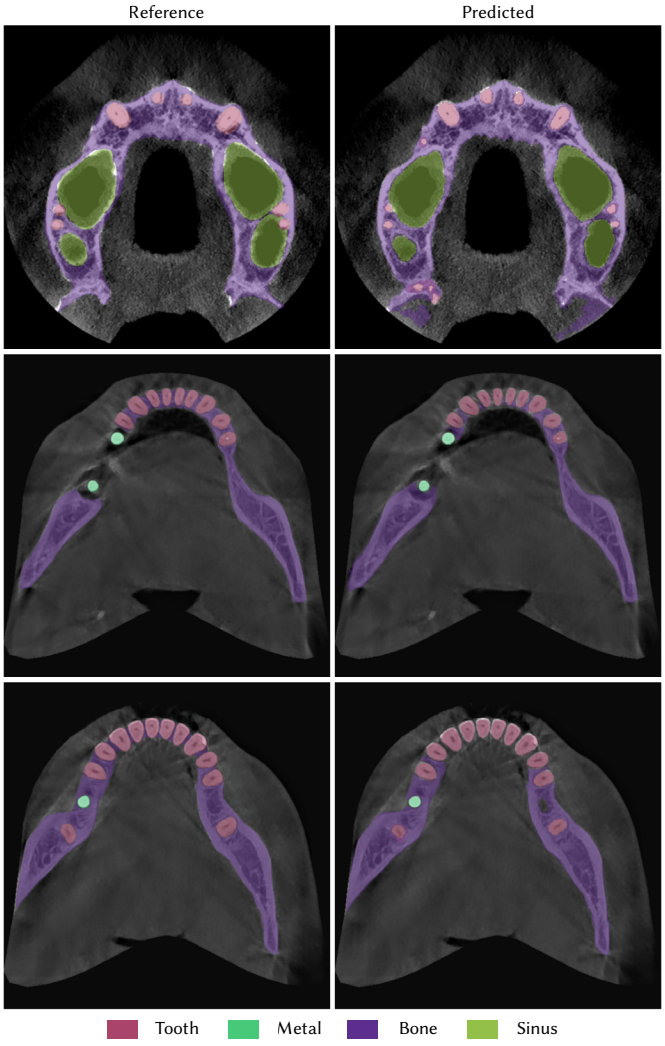



Fig. 4. Comparison of semantic segmentation maps of ground truth validation images with the predictions produced by the model.

## Acknowledgment

This research Project has been funded by the Comunidad de Madrid through the call Research Grants for Young Investigators from Universidad Politécnica de Madrid within the project NanoMoDL with grant number APOYO-JOVENES-21-MJK8FF-90-8I2C6U, and by the ITI Research Grant within the project "Bridging Centers, Enhancing Knowledge: A Research Protocol on Federated Learning for Dental Implant Classification and Pathology Identification in Periapical Radiographs." With grant number 1868-2024.

## References

[1] F. A. Yalda, J. Holroyd, M. Islam, C. Theodorakou, and K. Horner, "Current practice in the use of cone beam computed tomography: a survey of UK dental practices," *British Dental Journal*, vol. 226, no. 2, pp. 115–124, Jan. 2019, doi: https://doi.org/10.1038/sj.bdj.2019.49.

[2] A. J. Pakchoian DDS, "The Use of Cone Beam in Private Dental Practices in the United States: Cost and Reporting Patterns," Master's Thesis, University of Connecticut, 2016.

[3] L. Lenchik *et al.*, "Automated Segmentation of Tissues Using CT and MRI: A Systematic Review," *Academic Radiology*, vol. 26, no. 12, pp. 1695–1706, Dec. 2019, doi: https://doi.org/10.1016/j.acra.2019.07.006.

[4] N. O'Mahony *et al.*, "Deep Learning vs. Traditional Computer Vision," in *Advances in Computer Vision*, vol. 943, K. Arai and S. Kapoor, Eds., in Advances in Intelligent Systems and Computing, vol. 943. , Cham: Springer International Publishing, 2020, pp. 128–144. doi: https://doi.org/10.1007/978-3-030-17795-9_10.

[5] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022, doi: https://doi.org/10.1109/TNNLS.2021.3084827.

[6] H. Lamecker *et al.*, "Automatic segmentation of mandibles in low-dose CT-data," *International Journal of Computer Assisted Radiology and Surgery*, vol. 1, p. 393, 2006.

[7] D. Kainmueller, H. Lamecker, H. Seim, M. Zinser, and S. Zachow, "Automatic Extraction of Mandibular Nerve and Bone from Cone-Beam CT Data," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 76–83.

[8] H. Gao and O. Chae, "Individual tooth segmentation from CT images using level set method with shape and intensity prior," *Pattern Recognition*, vol. 43, no. 7, pp. 2406–2417, 2010, doi: https://doi.org/10.1016/j.patcog.2010.01.010.

[9] N. T. Duy, H. Lamecker, D. Kainmueller, and S. Zachow, "Automatic Detection and Classification of Teeth in CT Data," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 609–616.

[10] H.-T. Yau, T.-J. Yang, and Y.-C. Chen, "Tooth model reconstruction based upon data fusion for orthodontic treatment simulation," *Computers in Biology and Medicine*, vol. 48, pp. 8–16, 2014, doi: https://doi.org/10.1016/j.compbiomed.2014.02.001.

[11] Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao, "Tooth and Alveolar Bone Segmentation From Dental Computed Tomography Images," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 196–204, 2018, doi: https://doi.org/10.1109/JBHI.2017.2709406.

[12] Y. Fan *et al.*, "Marker-based watershed transform method for fully automatic mandibular segmentation from CBCT images," *Dentomaxillofacial Radiology*, vol. 48, no. 2, p. 20180261, 2019, doi: https://doi.org/10.1259/dmfr.20180261.

[13] D. X. Ji, S. H. Ong, and K. W. C. Foong, "A level-set based approach for anterior teeth segmentation in cone beam computed tomography images," *Computers in Biology and Medicine*, vol. 50, pp. 116–128, 2014, doi: https://doi.org/10.1016/j.compbiomed.2014.04.006.

[14] L. Hiew, S. Ong, K. W. Foong, and C. Weng, "Tooth segmentation from cone-beam CT using graph cut," in *Proceedings of the Second APSIPA Annual Summit and Conference*, ASC, Singapore, 2010, pp. 272–275.

[15] J. Keustermans, D. Vandermeulen, and P. Suetens, "Integrating Statistical Shape Models into a Graph Cut Framework for Tooth Segmentation," in *Machine Learning in Medical Imaging*, F. Wang, D. Shen, P. Yan, and K. Suzuki, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 242–249.

[16] P. Mortaheb, M. Rezaeian, and H. Soltanian-Zadeh, "Automatic dental CT image segmentation using mean shift algorithm," in *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*, 2013, pp. 121–126. doi: https://doi.org/10.1109/IranianMVIP.2013.6779962.

[17] Z. Cui, C. Li, and W. Wang, "ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Beam CT Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and*

*Pattern Recognition (CVPR)*, Jun. 2019. doi: https://doi.org/10.1109/CVPR.2019.00653.

[18] M. A. A. Hegazy, M. H. Cho, M. H. Cho, and S. Y. Lee, "U-net based metal segmentation on projection domain for metal artifact reduction in dental CT," *Biomedical Engineering Letters*, vol. 9, no. 3, pp. 375–385, Aug. 2019, doi: https://doi.org/10.1007/s13534-019-00110-2.

[19] S. Lee, S. Woo, J. Yu, J. Seo, J. Lee, and C. Lee, "Automated CNN-Based Tooth Segmentation in Cone-Beam CT for Dental Implant Planning," *IEEE Access*, vol. 8, pp. 50507–50518, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2975826.

[20] Y. Rao, Y. Wang, F. Meng, J. Pu, J. Sun, and Q. Wang, "A Symmetric Fully Convolutional Residual Network With DCRF for Accurate Tooth Segmentation," *IEEE Access*, vol. 8, pp. 92028–92038, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2994592.

[21] Y. Chen *et al.*, "Automatic Segmentation of Individual Tooth in Dental CBCT Images From Tooth Surface Map by a Multi-Task FCN," *IEEE Access*, vol. 8, pp. 97296–97309, 2020, doi: https://doi.org/10.1109/ACCESS.2020.2991799.

[22] T. J. Jang, K. C. Kim, H. C. Cho, and J. K. Seo, "A fully automated method for 3D individual tooth identification and segmentation in dental CBCT," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021, doi: https://doi.org/10.1109/TPAMI.2021.3086072.

[23] H. Wang, J. Minnema, K. J. Batenburg, T. Forouzanfar, F. J. Hu, and G. Wu, "Multiclass CBCT Image Segmentation for Orthodontics with Deep Learning," *Journal of Dental Research*, vol. 100, no. 9, pp. 943–949, 2021, doi: https://doi.org/10.1177/00220345211005338.

[24] Z. Zheng, H. Yan, F. C. Setzer, K. J. Shi, M. Mupparapu, and J. Li, "Anatomically Constrained Deep Learning for Automating Dental CBCT Segmentation and Lesion Detection," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 603–614, 2021, doi: https://doi.org/10.1109/TASE.2020.3025871.

[25] N. Morgan, A. Van Gerven, A. Smolders, K. de Faria Vasconcelos, H. Willems, and R. Jacobs, "Convolutional neural network for automatic maxillary sinus segmentation on cone-beam computed tomographic images," *Scientific Reports*, vol. 12, no. 1, p. 7523, May 2022, doi: https://doi.org/10.1038/s41598-022-11483-3.

[26] Z. Cui *et al.*, "A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images," *Nature Communications*, vol. 13, no. 1, p. 2096, Apr. 2022, doi: https://doi.org/10.1038/s41467-022-29637-2.

[27] G. Dot, A. Chaurasia, G. Dubois, C. Savoldelli, S. Haghighat, S. Azimian, A. R. Taramsari, G. Sivaramakrishnan, J. Issa, A. Dubey, T. Schouman, y L. Gajny, "DentalSegmentator: Robust open source deep learning-based CT and CBCT image segmentation," *Journal of Dentistry*, vol. 147, p. 105130, Aug. 2024, doi: https://doi.org/10.1016/j.jdent.2024.105130.

[28] F. Hu, Z. Chen, and F. Wu, "A novel difficult-to-segment samples focusing network for oral CBCT image segmentation," *Scientific Reports*, vol. 14, no. 1, p. 5068, Mar. 2024, doi: https://doi.org/10.1038/s41598-024-55522-7.

[29] Y. Jing, J. Liu, W. Liu, Z. Yang, Z. Zhou, and Z. Yu, "USCT: Uncertainty-regularized symmetric consistency learning for semi-supervised teeth segmentation in CBCT," *Biomedical Signal Processing and Control*, vol. 91, p. 106032, May 2024, doi: https://doi.org/10.1016/j.bspc.2024.106032.

[30] F. Nogueira-Reis, N. Morgan, I. R. Suryani, C. P. M. Tabchoury, and R. Jacobs, "Full virtual patient generated by artificial intelligence-driven integrated segmentation of craniomaxillofacial structures from CBCT images," *Journal of Dentistry*, vol. 141, p. 104829, Feb. 2024, doi: https://doi.org/10.1016/j.jdent.2023.104829.

[31] C. Wang, J. Yang, B. Wu, R. Liu, and P. Yu, "Trans-VNet: Transformer-based tooth semantic segmentation in CBCT images," *Biomedical Signal Processing and Control*, vol. 97, p. 106666, Nov. 2024, doi: https://doi.org/10.1016/j.bspc.2024.106666.

[32] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, Sep. 2022, doi: https://doi.org/10.1145/3505244.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, y N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint, arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929, doi: https://doi.org/10.48550/arXiv.2010.11929.

[34] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, y C. L. Zitnick, "Microsoft COCO: Common Objects in Context," en *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, y T. Tuytelaars, Eds. Cham, Switzerland: Springer International Publishing, 2014, pp. 740–755.

[36] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[37] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing through ADE20K Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 5122–5130. doi: https://doi.org/10.1109/CVPR.2017.544.

[38] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 5000–5009. doi: https://doi.org/10.1109/ICCV.2017.534.

[39] L. Wang, Y. Gao, F. Shi, G. Li, K.-C. Chen, Z. Tang, J. J. Xia, and D. Shen, "Automated segmentation of dental CBCT image with prior-guided sequential random forests: Automated segmentation of dental CBCT image," *Medical Physics*, vol. 43, no. 1, pp. 336–346, Dec. 2015, doi: https://doi.org/10.1118/1.4938267.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241.

[41] D. M. Pelt and J. A. Sethian, "A mixed-scale dense convolutional neural network for image analysis," *Proceedings of the National Academy of Sciences*, vol. 115, no. 2, pp. 254–259, 2018, doi: https://doi.org/10.1073/pnas.1715832114.

[42] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2014, *arXiv*. doi: https://doi.org/10.48550/ARXIV.1409.0473.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[45] E. A. Nadaraya, "On Estimating Regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964, doi: https://doi.org/10.1137/1109020.

[46] G. S. Watson, "Smooth Regression Analysis," *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 26, no. 4, pp. 359–372, 1964.

[47] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

[48] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint, arXiv:1607.06450*, 2016. [Online]. Available: https://arxiv.org/abs/1607.06450.

[49] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, and D. Liang, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 10012–10022.

[51] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7478–7498, June 2024, doi: https://doi.org/10.1109/TNNLS.2022.3227717.

[52] J. Brooks, "COCO Annotator." 2019. [Online]. Available: https://github.com/jsbroks/coco-annotator/

[53] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," in *Neural Information Processing Systems*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235829267

[54] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 936–944. doi: https://doi.org/10.1109/CVPR.2017.106.

[55] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. S. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, and Z. Lu, Eds., Cham: Springer International Publishing, 2017, pp. 240–248.

[56] F. Bolelli, S. Allegretti, L. Baraldi, and C. Grana, "Spaghetti Labeling: Directed Acyclic Graphs for Block-Based Connected Components Labeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 1999–2012, 2020, doi: https://doi.org/10.1109/TIP.2019.2946979.

[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[58] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021, doi: https://doi.org/10.1109/TPAMI.2021.3059968.

Rafael C. Giménez-Aguilar

Rafael C. Giménez-Aguilar is a Ph.D. candidate and researcher at the Universidad Politécnica de Madrid. He has a background in Biomedical Engineering, initiating his career as an assistant researcher on advanced radiology at the Carlos III University of Madrid (2020), where he authored and co-authored three conference publications. He then pursued an M.Sc. in Artificial Intelligence (2021) at his current university, enrolling as a researcher at its Biomedical Informatics Group, where he focuses on projects employing AI on Computer Vision and NLP tasks in a Biomedical context.

Sergio Paraíso-Medina

Sergio Paraíso-Medina is a PhD in Artificial Intelligence from the Universidad Politécnica de Madrid (UPM). Since 2018 he is teaching at the Departament of "Lenguajes Sistemas Informáticos e Ingeniería de Software", ETSI Informáticos at UPM, where he is currently a Teaching Assistant, where he also belongs to the Biomedical Informatics Group (www.gib.dia.fi.upm.es) since 2010. He has participated in up to 9 European and Spanish research projects, highlighting the work carried out in the European INTEGRATE and EURECA projects and in the national health projects CIMED and CICLOGEN, where the researcher was responsible for the integration and interoperability of clinical data and has published most of his articles.

Miguel García-Remesal

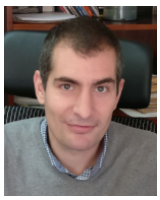
Miguel García-Remesal received a Master's degree in Computer Science from the Universidad Politécnica de Madrid (Madrid, Spain) in 2001 and was awarded a Ph.D. in Computer Science from the same university in 2006. He was a visiting scholar at the Department of Computer Science of Rutgers University (NJ, USA) in 2005 and the Department of Biomedical Informatics of the University of Utah (UT, USA) in 2010. He is currently an associate Professor (tenured) with the Department of Artificial Intelligence of UPM, Director of the Ph.D. program of Artificial Intelligence from UPM, and a senior researcher of the Biomedical Informatics group at the same university. His research interests include Biomedical Informatics, Bioinformatics, Text Analytics, Information Retrieval, and Heterogeneous Database Integration.

Guillermo Jesús Pradíes-Ramiro

Guillermo Jesús Pradíes-Ramiro is a full-time professor and Head of the Prosthodontics Department at the Complutense University of Madrid (UCM). Currently he is the PRESIDENT of the Spanish Society of Prosthodontics and Esthetic Dentistry. He obtained his master's degree in Prosthetic Dentistry at the University Complutense in 1994 and has accreditation as a Specialist in Prosthodontics by the European Prosthodontics Association. His research activity is focused on three main topics: Artificial Intelligence in Dentistry, CAD/CAM digital workflows, and aesthetic materials evaluation for prosthodontics uses. He has published more than 90 scientific articles and various textbooks. He is an international Lecturer in Prosthetic and Implant Dentistry and has won various national and international clinical and research awards. He is the director of the master's degree Program in "Restorative Dentistry based on New Technologies" at the UCM as well as Associate Editor of Brazilian Dental Science Journal and Assistant editor of European Journal of Prosthodontics and Restorative Dentistry. He is also a member of the council and the specialist committee of the European Prosthodontic Association, organization from which he also was President in 2018.

Monica Bonfanti Gris

Monica Bonfanti Gris is a Ph.D. candidate specializing in the application of AI in dental diagnostics. She initiated her research journey in 2020, earning her Master of Science degree. Monica furthered her studies through the master's degree program in Restorative Dentistry with a focus on New Technologies at Complutense University of Madrid. Despite her young age, she has already published articles in reputable journals and actively participates in dental research. Monica's work centers on the integration of AI into dental diagnostics, a field with significant potential for improving patient care and outcomes. Her dedication to this emerging area promises to drive important advancements in dental healthcare.



Raul Alonso-Calvo

Raul Alonso-Calvo is a PhD in Computer Science from the Universidad Politécnica de Madrid (UPM). He has been a visitor researcher at Universidade de Aveiro DETI-IEETA (Portugal) (2010), and Oxford University (UK) (2014). Since 2010 he has been at the Departamento de Lenguajes Sistemas Informáticos e Ingeniería de Software, ETSI Informáticos at UPM, where he is currently an Associate Professor. He has been a member of the Biomedical Informatics Group at UPM since 2001. His research interests are mainly focused on clinical research informatics, biomedical interoperability standards, database integration and preprocessing, biomedical image processing and information retrieval in biomedicine. He has been the author and co-author of research papers in several journals. He has participated in EU research projects since 2001 and in recent years he was involved in INTEGRATE: Driving Excellence in Integrative Cancer and EURECA: Enabling information re-use by linking clinical Research and Care.