IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 April 2024, accepted 19 May 2024, date of publication 3 June 2024, date of current version 11 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3408629

RESEARCH ARTICLE

Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal

LUCA LUMETTI, VITTORIO PIPOLI, FEDERICO BOLELLI, (Associate Member, IEEE), ELISA FICARRA, AND COSTANTINO GRANA

Dipartimento di Ingegneria "Enzo Ferrari," Università degli Studi di Modena e Reggio Emilia, 41125 Modena, Italy

Corresponding author: Federico Bolelli (federico.bolelli@unimore.it)

This work was supported in part by the Department of Engineering "Enzo Ferrari" of the University of Modena through the Fondo di Ateneo per la Ricerca 2023 (FARD-2023) and in part by European Union's Horizon 2020 Research and Innovation Program under Grant 965193.

ABSTRACT Segmentation of the Inferior Alveolar Canal (IAC) is a critical aspect of dentistry and maxillofacial imaging, garnering considerable attention in recent research endeavors. Deep learning techniques have shown promising results in this domain, yet their efficacy is still significantly hindered by the limited availability of 3D maxillofacial datasets. An inherent challenge is posed by the size of input volumes, which necessitates a patch-based processing approach that compromises the neural network performance due to the absence of global contextual information. This study introduces a novel approach that harnesses the spatial information within the extracted patches and incorporates it into a Transformer architecture, thereby enhancing the segmentation process through the use of prior knowledge about the patch location. Our method significantly improves the Dice score by a factor of 4 points, with respect to the previous work proposed by Cipriano et al., while also reducing the training steps required by the entire pipeline. By integrating spatial information and leveraging the power of Transformer architectures, this research not only advances the accuracy of IAC segmentation, but also streamlines the training process, offering a promising direction for improving dental and maxillofacial image analysis.

INDEX TERMS CBCT, inferior alveolar canal, medical imaging, 3D imaging, transformers, patch-based learning.

I. INTRODUCTION

Maxillofacial surgery represents a complex challenge due to the presence of the Inferior Alveolar Nerve (IAN). Among bone canals containing blood vessels and nerves, the one containing the IAN, known as the Inferior Alveolar Canal (IAC) or simply Mandibular Canal (MC), supplies sensation to the lower teeth, lips, and chin, and its position (Fig. 1) must be carefully identified before surgical intervention. Avoiding contact with the IAN is a priority during implant placement, molar extraction, and many other craniofacial procedures to prevent aches, pain, and temporary or permanent paralysis [1]. To achieve such a goal, the preoperative treatment planning and simulation requires a strong and accurate IAC segmentation [2], [3] that is usually performed upon data acquired with Cone Beam Computer Tomography (CBCT), a low-radiation and cheap 3D image modality.

Unfortunately, achieving meticulous 3D annotations is a time-intensive task that requires the eyes of expert clinicians. Consequently, CBCTs are usually condensed into 2D panoramic views providing an approximation of relevant information and used for canal segmentation or other preoperative planning. This approach, known as panoramic radiography, prevents from determining the 3D rendering of the entire canal and the connected anatomical structures.

Recent advancements in the field of Deep Learning have significantly impacted the medical imaging domain, particularly through methods based on Convolutional Neural Networks (CNNs) [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. While CNNs excel across a spectrum of computer vision tasks, the rise of attention mechanisms and Transformer methodologies has propelled them to prominence, surpassing

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

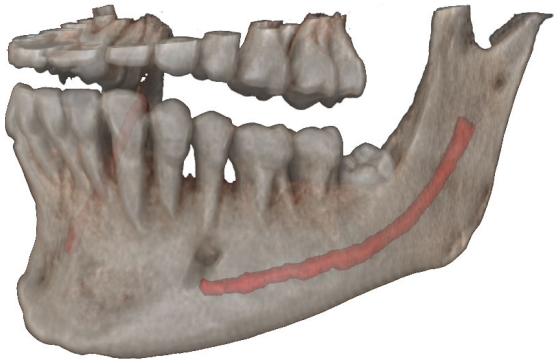


FIGURE 1. CBCT scan with the inferior alveolar canal marked in red.

CNNs in various domains, including medical imaging [14], [15]. Initially introduced by Vaswani et al. [16] for machine translation, Transformers have evolved into state-of-the-art solution for numerous Natural Language Processing (NLP) tasks. Subsequently, the advent of the Vision Transformer (ViT) architecture marked a turning point, extending Transformers' dominance also into computer vision and, as a consequence, diverse domains like image segmentation. Notably, Transformers have recently been incorporated into U-Net architectures [17] for both 2D and 3D medical image segmentation, demonstrating encouraging outcomes [14], [15]. Motivated by the flexibility and potential of Transformer architectures, we propose to integrate this methodology into our model to address the 3D IAC segmentation challenge.

Independently from the technique employed, when segmenting large 3D volumes, it is often required to adopt a patch-based processing approach to reduce input data size and fit memory constraints. This makes the training feasible but compromises the neural network performance due to the absence of global contextual information. Hence, this paper introduces a novel approach called Memory Augmented Transformer with Absolute Positional information, MATAP, a 3D architecture enhanced by a memory-augmented Transformer encoder that effectively harnesses absolute spatial coordinates, mitigating the challenges associated with patch-based training.

More specifically, our proposal builds from the standard 3D U-Net architecture introducing a memory-augmented Transformer in the bottleneck. By capitalizing on the inherent capacity of Transformers to model interactions between all pairs of elements within a given sequence, we aim to enhance the flow of information among the elements of the U-Net bottleneck, thus increasing contextualization. Moreover, we leverage such a flow of information to effectively inject contextual information related to the processed patches, i.e., their position within the entire volume, mitigating the patch-based learning-related issues. The ''memory'' is an additional refinement that supports the model to retain crucial prior concepts that may be challenging to be directly extracted from image features, but are nonetheless valuable for interpretation.

The key contributions of this research can be summarized as follows:

- Design of a memory-augmented Transformer module capable of leveraging absolute spatial coordinates when dealing with patch-based learning;
- Introduction of an innovative deep learning architecture tailored for 3D IAC segmentation, effectively addressing the limitations of patch-based training;
- Release of the source code,[1] enabling precise replication of all the proposed experiments and future comparison with additional proposals on the field.

The subsequent sections of this paper are organized as follows. Section II summarizes related works and Section III introduces and describes the dataset employed for evaluation. Our proposed methodology is then detailed in Section IV, followed by the presentation of experiment outcomes in Section V. The paper concludes with Section VI, providing a final summary and directions for future research.

II. RELATED WORKS

In the field of dentistry and maxillofacial surgery, three-dimensional (3D) imaging has emerged as a fundamental technology for accurate diagnosis [18]. Initial studies primarily focused on the utilization of Computed Tomography (CT) scans [19], [20]. However, with the introduction and widespread adoption of cone beam computed tomography in the early 2000s [21], significant attention has been directed towards the development of automated systems for the segmentation of the inferior alveolar canal. This research has encompassed both classical computer vision methods [22], [23], [24], [25], [26], [27], [28] and, more recently, machine learning and deep learning approaches [29], [30], [31], [32].

For what concerns classical computer vision methods, Kainmueller et al. [22] proposed a fully automatic approach based on a combined Statistical Shape Model of the bone surface and nerve course. Their method improved nerve position reconstruction using a Dijkstra-based tracing algorithm. Abdolali et al. [25] presented a similar solution with a pre-processing step based on low-rank decomposition and fast marching for optimal path determination between the mandibular and mental foramen. However, these methods require manual segmentation of the mandible bone in the training annotation, adding extra manual effort. Moris et al. [26] took a different approach by considering the volume from multiple perspectives and extracting the canal by searching for candidate positions with high similarity to an IAC ideal model. However, their method relies on predefined thresholds, often excluding parts of the canal due to low contrast in CBCT scans. Wei and Wang [27] introduced a method based on a curved Multi-Planar Reconstruction (MPR) image set and clustering of texture features to enhance image contrast. They segmented the mandibular canal edges using 2D line-tracking and fitting the results with a fourth-order polynomial.

[1]The source code is available at https://github.com/AImagelab-zip/alveolar_canal.

While classical computer vision approaches have made significant contributions, the most successful models for the segmentation of the IAC lie within the domain of machine learning and deep learning. Notably, Jaskari et al. [31] presented one of the pioneering applications of deep learning for mandibular canal segmentation. Their approach involved training a fully convolutional network using a dataset of coarsely annotated 3D scans. On average, each canal was annotated with 10 manually assigned control points, which were subsequently interpolated using a 3D spline and converted into volumetric representations by placing disks with a fixed diameter of 3.0 mm on the planes orthogonal to the spline. This deep learning approach demonstrated superior performance compared to previous methods relying on Statistical Shape Models. However, it encountered limitations due to the lack of finely annotated voxel-level data and the sub-optimal quality of segmentation masks generated automatically from coarse annotations.

Another take on the use of CNNs for mandibular canal segmentation was proposed by Kwak et al. [30]. Their work involved training 2D and 3D models based on the SegNet [33] and U-Net [17], [34] architectures using a privately annotated dataset and an arguable evaluation metric [35]. Additionally, Lahoud et al. [36] employed a standard 3D U-Net model for IAC segmentation. However, neither the dataset nor the experimental code from these studies is publicly available, limiting the possibility of direct comparisons with our work.

To address the challenges in inferior alveolar canal segmentation, Cipriano et al. [35] introduced a significant breakthrough by proposing the first publicly available dataset of 3D annotated CBCT scans of the human jaw, named Maxillo, alongside a state-of-the-art deep learning model for the 3D segmentation of the IAC called PosPadUNet3D. The Maxillo dataset comprises 347 CBCT scans, with 91 of them featuring 3D annotations meticulously generated by radiologic technologists using the IACAT tool [37]. This marks a substantial advancement in publicly accessible datasets for the segmentation of the inferior alveolar canal. The segmentation pipeline based on the PosPadUNet3D model used a three-step training procedure: in the initial step, known as "deep label expansion," the network was trained using CBCT volumes paired with their corresponding sparse 2D labels to generate dense 3D annotations. Next, they employed this network to generate synthetic 3D annotations for the 256 volumes for which only a 2D annotation is available. Subsequently, the synthetic 3D labels were employed for pre-training the segmentation network, which was further fine-tuned using 3D annotations performed by medical experts. Notably, unlike other 3D U-Net-based algorithms in the literature, PosPadUNet3D incorporates patch positional information in the bottleneck and employs padded convolutions to preserve tensor dimensionality.

After the appearance of the Maxillo dataset, different authors employed public data to evaluate their proposals, making them verifiable and comparable.

In [38], Usman et al. proposed a two-stage approach also based on the U-Net architecture. Their methodology was formulated on the hypothesis that the predominant challenge in segmenting the inferior alveolar canal relates to the class imbalance between the mandibular canal and the background. To address this issue, they initially apply a CNN to identify and isolate volume regions where the canal is likely to be located (Regions of Interest, ROIs), thereby substantially reducing background interference. Subsequently, in the second phase, they leverage U-Net architecture to perform the segmentation of the mandibular canal, exclusively within the ROIs.

Another contribution was by Zhao et al. [39] and, similarly to [38], it is based on a two-stage approach. The author proposed a whole mandibular canal segmentation using transformed dental CBCT volume in the Frenet frame. They first extracted the mandibular centerline via automatic segmentation of the mandible and localization of the mandibular foramen and mental foramen. The sub-volumes containing the mandibular canal information were then obtained using a double reflection method based on the Frenet frame. The transformed sub-volumes were fed into the 3D segmentation network (again U-Net-based), and the clDice loss was used to constrain the topology of the mandibular canal. Lastly, the prediction masks were inversely transformed back into the original CBCT images to obtain final segmentations.

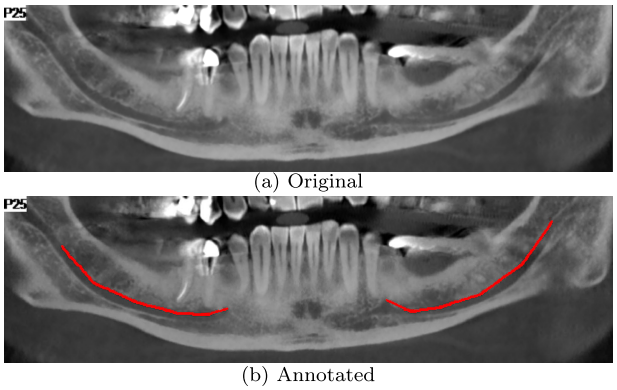
Liu et al. [40] introduced the Frequency-domain Attention Module (FAM) in the U-Net architecture. Although the proposed FAM only includes 56 additional parameters with respect to U-Net, they significantly improved the IAN canal segmentation accuracy.

III. DATASET

The maxillofacial dataset employed in our experiments, named *ToothFairy*, is an enhanced version of the *Maxillo* dataset proposed by Cipriano et al. [41]. It has been employed in the homonymous MICCAI 2023 challenge hosted on the *Grand Challenge* platform. This dataset was created using the updated version of the IACAT tool [37], specifically IACAT 2.0, developed in [42]. The innovative features introduced enhanced the clinicians' performance during the canal identification process and reduced the annotation time. Consequently, compared to the Maxillo dataset, ToothFairy improves both the quality and quantity of the 3D annotations, with an average increase of 61.9% in the number of annotated

TABLE 1. Summary description of the ToothFairy dataset.

Dataset	Split	Volumes	Annotations	
			2D	3D
Primary	Training	153	✓	✓
	Validation	8	✓	✓
	Testing	15	✓	✓
Secondary	Training	290	✓	✗



(a) Original

(b) Annotated

FIGURE 2. Example of a 2D annotation obtained from a panoramic view.

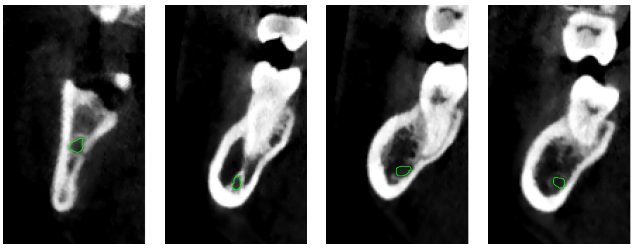


FIGURE 3. Examples of cross-sectional views and corresponding annotations. Combining the closed splines generated from different views will produce the final voxel-level dense annotation.

voxels. Thanks to the aforementioned tool, clinicians were able to carry out annotations at a faster rate, providing 62 additional 3D-annotated volumes, as well as improving 40 of the 91 3D-ground-truth annotations already available within the Maxillo dataset. More specifically, the total number of CBCT scans in the newer dataset increased from 347 to 443. For convenience, the details about ground-truth, test, and training sets are summarized in Tab. 1. A distinction between the so-called *primary* and *secondary* dataset, containing respectively 3D (dense) annotations or 2D (sparse) annotations, is provided in Sec. III-A.

In the same study by Lumetti et al. [42], in addition to the clinical validation performed by medical experts, the authors demonstrated that all tested deep learning models for IAC segmentation performed better with the ToothFairy dataset.

Therefore, the majority of the experiments carried out and described in the following of this article have been performed using the ToothFairy dataset.

All 3D CBCT volumes were acquired from the Affidea center located in Modena, Italy, a leading pan-European healthcare group specializing in advanced diagnostics, out-patient services, laboratory analyses, physiotherapy and rehabilitation, and cancer diagnosis and treatment. The organization operates 312 centers across 15 countries, with approximately 11,000 professionals.

The annotations have been performed by medical experts with more than five years of experience in the maxillofacial field. No multiple annotations are available for a specific patient, meaning that all the CBCT volumes are annotated by a single expert only.

The dataset is publicly available after user registration and it can be downloaded from https://ditto.ing.unimore.it/toothfairy/. Such availability, combined with the public release of the source code allows for a complete reproducibility of our experiments and empiric verification of our claims.

A. 2D AND 3D ANNOTATIONS

The diagnostic technicians responsible for the examinations were also involved in the initial annotation process of the mandibular canal, providing what we refer to as ''sparse annotations.'' These are performed on 2D panoramic views of the jawbone and are routinely used in surgical practice to measure the height and depth of implant placement sites, thereby avoiding injuries to the inferior alveolar nerve.

In these sparse labels, the upper boundary of the canal is marked along the entire dental arch, offering a useful approximation of the nerve position. In this context, the annotation process begins with the selection of an axial slice from the original volume. The central position of the jawbone is roughly identified upon this slice with the so-called panoramic base curve, i.e., a line employed to generate the panoramic view (Fig. 2a). This view is composed of the voxels lying on the curved surface identified by the base curve and orthogonal to the axial plane. The inferior alveolar canal should be distinctly identifiable within this panoramic view. An example of a 2D annotation provided by an expert technician is shown in Fig. 2b. Both the *primary* and *secondary* datasets contain sparse annotation.

Instead, the 3D annotation process is performed using the IACAT tool [42]. This makes use of an automatically computed 2D base curve to propose Cross-Sectional Views (CSVs) where radiologic technologists can annotate the inferior alveolar canal by drawing a closed Catmull-Rom spline (Fig. 3). As one can imagine, this operation must be conducted for a significant number of CSVs to ensure sufficient accuracy in the annotation. The number of proposed CSVs depends on a configurable parameter that determines the distance between two consecutive views.

After the annotation process, the closed splines produced by medical experts are saved as the coordinates of their control points. Subsequently, the final ground-truth volume, which is both smooth and precise, is generated from this set of control points using the α-shape algorithm [43], as described in [35]. Dense annotations are only available for the *primary* dataset.

When comparing the two procedures, it becomes evident that obtaining sparse annotations is a quick and straightforward process, while creating dense labels from 3D volumes is significantly more laborious and time-consuming. Consequently, researchers often have limited access to densely annotated volumes and typically reserve them for the test set.

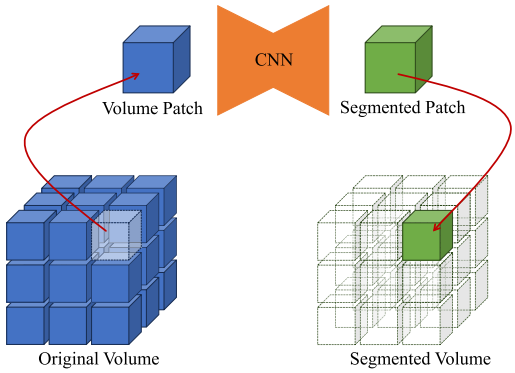


FIGURE 4. Visualization of a patch-based training. From the original volume, depicted in blue, a sub-volume is extracted and fed into the network. The output, depicted in green, is then placed in the same position where the patch has been extracted.

IV. METHODS

A. PATCH-BASED LEARNING

In recent years, learning using patch-based representations has become increasingly popular, especially when dealing with visual tasks. There are specific scenarios that make patch-based learning the only viable approach, e.g., when targeting complex, high-dimensional inputs, or when the computational resources available are limited or should be kept so. The segmentation or classification in whole-slide images, as well as the segmentation of anatomical structures in 3D volumes, are noticeable medical imaging applications requiring such a kind of learning procedure. Indeed, feeding a neural network with gigapixel images or hundreds of millions of voxels coming from 3D volumes is not a feasible approach.

To provide the reader with an example, numerous popular image classification models employ an input image size of $3 \times 224 \times 224$, equivalent to $1.5 * 10^5$ pixels [6], [7], [8]. Conversely, the CBCT scans of the aforementioned ToothFairy dataset have a spatial dimension of $169 \times 342 \times 370$ voxels, for a total of $2.1 * 10^7$ voxels. To meet memory constraints, the simple downsampling of the input data is counterproductive whenever the preservation of fine-grained details is crucial. The increase in the spatial dimensions propagates in every layer of the neural network, making it often impossible to process even a single volume through a GPU. To overcome this limitation, it is common in medical imaging to train neural networks using subsets extracted from the original data [29], [30], [31], [44], as depicted in Fig. 4.

Such an approach mitigates memory constraints, but it also introduces additional hurdles that must be taken into account: the loss of global information due to restricted, patch-limited receptive fields, ambiguity in segmenting objects situated at the intersections of multiple patches, and potential artifacts arising from the boundaries of these patches. Furthermore, these challenges become particularly prominent when the object to be segmented is small in comparison to the entire volume: the segmentation of the IAC is an example of such circumstances.

A first proposal to overcome the patch-based learning drawbacks in the segmentation of the IAN is introduced by Cipriano et al. [35] with the PosPadUNet3D. The authors suggested leveraging the positional information from the coordinates of extracted patches by simply projecting and concatenating these coordinates within the network bottleneck. Although this approach demonstrated some improvements in the network performance, the observed enhancement was limited, and the aforementioned major issues still persisted.

B. THE PROPOSED APPROACH

With this paper, we introduce a novel deep-learning model designed for the segmentation of 3D scans of the inferior alveolar canal. Our proposed approach extends the existing PosPadUNet3D [35] architecture by addressing the limitations associated with patch-based learning through the utilization of Transformers capable of exploiting contextual information.

Our model (Fig. 5) incorporates memory-augmented Transformer encoder blocks. By capitalizing on the inherent capacity of Transformers to model interactions between all pairs of elements within a given sequence, we aim to enhance the flow of information among the elements of the U-Net bottleneck. Moreover, we leverage this to effectively inject contextual information related to the processed patches.

In practice, we introduce a specialized token that captures the absolute position of the patch within the original volume, referred to as `[ABS]`. This is accomplished by projecting the 3D coordinates of two opposite corners of the patch into the embedding of the bottleneck, exploiting a learnable matrix of dimension $6 \times d_{model}$. Subsequently, we concatenate this token with the remaining elements of the bottleneck, allowing its information to influence their representations through the Transformer encoder.

It is worth noticing that Transformers already employ positional encoding to describe the location of a token in a sequence. Such an encoding provides information about the position of (groups of) voxels within the current patch only. Instead, our `[ABS]` token encodes the position of a patch with respect to the entire volume. However, the inbuilt positional encoding of the Transformer architecture must not be applied to the `[ABS]` because all the other tokens should be able to employ its information independently from their position. To achieve this goal the positional encoding is applied before concatenating any other element. Again, this ensures that the `[ABS]` token remains positionally untied from the rest of the sequence. This disentangled approach allows each element to pay attention to the special token's information and vice versa, regardless of its position in the sequence.

Additionally, inspired by the advancements in the vision-language domain, we incorporate a memory-augmented Transformer encoder within our architecture. The integration of Transformer memory has demonstrated considerable effectiveness in tasks such as image captioning [45]. This mechanism enables the Transformer to retain crucial prior

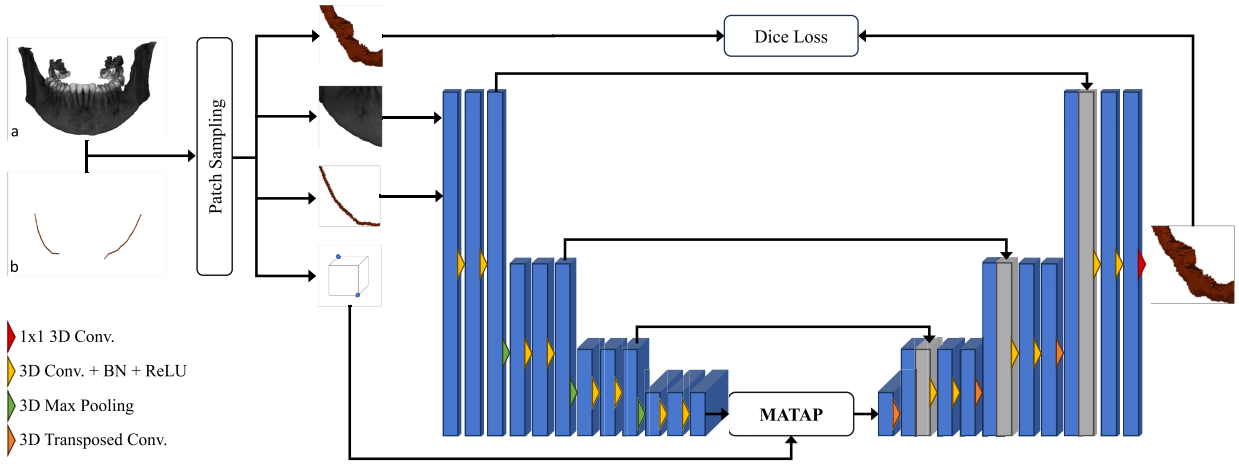


FIGURE 5. Proposed architecture with the MATAP module. (a) is the input volume and (b) is the corresponding sparse annotation, which is only concatenated to (b) during the generation phase, while it is not employed for the segmentation phase. The detailed visualization of the MATAP module is reported in Fig. 6.

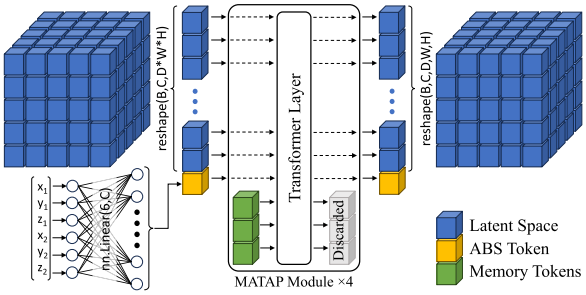


FIGURE 6. Graphical representation of the proposed module. The letters `B`, `C`, `D`, `W`, and `H` represent respectively batch size, channels, depth, height, and width. The patch coordinates `[x1, y1, z1, x2, y2, z2]` are projected using a linear layer producing the `[ABS]` token represented in yellow. The activation map obtained in the bottleneck of U-Net before the first transposed convolution is flattened across the spatial dimension, concatenated with the `[ABS]` token and `M×4` memory tokens, which are different for each transformer layer. After the processing of the 4-layer transformer, the `[ABS]` and the memory tokens are removed and the remaining output is reshaped back to the original spatial dimensionality.

concepts that may be challenging to be directly extracted from image features, but are nonetheless valuable for interpretation. Recognizing the applicability of this approach to the patch-based learning paradigm, wherein each patch is extracted from a wider context, we harness the power of Transformer memory to incorporate external information thus enhancing the processing of individual patches. In practice, memory tokens are learnable vector representations that are concatenated to the input of each transformer encoder layer and removed after being processed. A graphical summary is provided in Fig. 6.

C. FILTERING WITH THE HANN WINDOW FUNCTION

Even if the proposed memory-augmented Transformer-based encoder with `[ABS]` token mitigates the lack of global information in patch-based learning and reduces the segmentation ambiguity of objects situated on patch borders, we still need to deal with noise and artifacts generated at patch boundaries.

Taking inspiration from the field of audio encoding [46], we introduced a post-processing algorithm based on the Hann windows function to tackle the presence of artifacts near patch edges. The Hann window function is defined as:

$$W_{\text{Hann}}(i) = \frac{1}{2}\left(1 - \cos\frac{2\pi i}{I}\right) \tag{1}$$

where i is an element in the considered interval I. This function is symmetric, peaking at 1 in the middle of the window and tapering to 0 at the edges. An intriguing property of this function is that the sum of two Hann windows, each shifted by $\frac{I}{2}$ (50%), is equivalent to a rectangular window of width I and height 1:

$$W_{\text{Hann}}(i) + W_{\text{Hann}}\left(i + \frac{I}{2}\right) = 1 \tag{2}$$

Such a property is exploited in audio encoding to eliminate border artifacts. This is achieved by multiplying the Hann window with frames that overlap by 50%, before summing them together. While this approach is defined in 1D for audio, it can be extended to multiple dimensions, making it applicable to 3D images:

$$W_{\text{Hann}}(i,j,k) = W_{\text{Hann}}(i)W_{\text{Hann}}(j)W_{\text{Hann}}(k) \tag{3}$$

where i, j, k identify a point in the space. When implementing this filtering, particular care should be exercised to ensure that the window function applied in the 3D space still sums to one also on volume borders.

By applying the proposed 3D extension of the Hann filtering to the output segmented patches produced by our model, we are able to reduce the aforementioned noise on patch borders (Fig. 7) and improve the overall performance (Sec. V).

D. MODEL TRAINING

The adopted model training procedure partially follows the one described in Cipriano et al. [35]. In particular, we use a

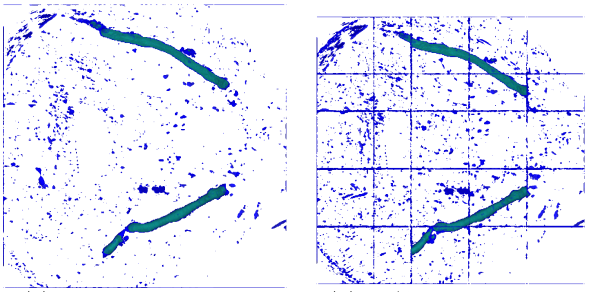


FIGURE 7. (a) is an axial plane extracted from a predicted volume after applying the Hann windows function, (b) is the same plane obtained without any post-processing. In both images, blue represents logits that have a value higher than 10^{-4}. The post-processing significantly reduces artifacts appearing close to patch borders. Even if most of these artifacts do not cause any issues, the ones which are close to the IAC badly influence the final segmentation.

two-step procedure composed of an initial step called "deep label expansion" or "generation phase" and a second one that consists of a standard segmentation training. In the "deep label expansion," the network is trained using CBCT volumes paired with their corresponding sparse 2D labels to generate 3D dense annotations. The rationale behind this operation is to obtain a model that can leverage the sparse 2D labels (that are available for both the primary and the secondary ToothFairy datasets) to create dense synthetic 3D annotations when they are not available. The second step consists of merging the initial training set provided with true labels (primary) with the synthetically annotated CBCT volumes (generated from the secondary dataset). Thus, a total of 443 3D annotated volumes is obtained as a train set. Still, 8 and 15 volumes from the primary dataset are available for the validation and test set respectively.

Using the above-mentioned set of data, our MATAP segmentation model is trained to output 3D masks representing the inferior alveolar canal, and consequently evaluated.

Differently from the procedure proposed by Cipriano et al. [35], we employ human and synthetically generated 3D annotations in a single training stage, without distinguishing between pre-training, performed by Cipriano et al. with only synthetically generated annotations, and fine-tuning, performed in [35] by means of true labels only. Our improved generation step produces synthetic annotations that are very close to the ground truth, even more than those obtained by Cipriano et al. [35]. Hence, we can save time and simplify the process by relying on a single-phase training procedure, without degrading the overall segmentation performance.

E. EVALUATION PROTOCOL

Considering the stochastic nature of parameter fitting in neural networks, different models trained under identical experimental conditions yield slightly varying outputs. These variations are usually approximated by a normal distribution. Taking this into consideration, to ensure the robustness and reliability of our proposal, we run each experiment ten times, resulting in a population size, N, of ten: $X_1, \ldots, X_{10}$. Each experiment involved training the model with the same experimental conditions, but with a different random seed employed for the initialization. The corresponding output values of the test evaluation metric, specifically the Dice, are recorded obtaining a population of test metrics.

To validate the strength of our proposal and compare the above-mentioned populations, we employed different standard statistical tools, which are detailed in the following.

1) CONFIDENCE INTERVAL

A statistical tool for assessing the potential range around the estimate of a statistical measure for a given population is the Confidence Interval (CI), which also allows to highlight how stable an assessment is.

To compute the CIs, we advocate for the Student's t-distribution with $N-1$ degrees of freedom. The strength of the t-distribution comes from its ability to adjust for smaller sample sizes (and therefore fewer degrees of freedom) by effectively having a more conservative estimate of probability density with respect to the normal distribution. Having a relatively small population of 10 samples, the t-distribution is well-suited for our purpose, allowing us to determine the range of plausible values around the estimated mean.

Practically, for each experiment, we calculate the standard deviation (S) using the *unbiased estimator* as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{N}(\bar{X} - X_i)^2}{N-1}} \tag{4}$$

where N is the number of data points, X_i are the observed values and $\bar{X} = \frac{1}{n}\sum_i X_i$ is their average. Then, the mean (Mean), lower bound (LB), and upper bound (UB) values of the CIs are computed as follows:

$$Mean = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{5}$$

$$LB = \bar{X} - t_{\alpha/2,\nu}\frac{S}{\sqrt{N}} \tag{6}$$

$$UB = \bar{X} + t_{\alpha/2,\nu}\frac{S}{\sqrt{N}} \tag{7}$$

where, again, N is the number of data points and X_i are observed values. In this context, α represents the significance level (in our case equal to 0.05, so that the confidence level is 0.95), ν is the number of degrees of freedom, and $t_{\alpha/2,\nu}$ is the t-distribution evaluated at $\alpha/2$, ν.

2) HYPOTHESIS TESTING

It involves formulating two competing hypotheses, the *null hypothesis* (H_0) and the *alternative hypothesis* (H_A), and then collecting data to assess if there is enough evidence in a sample data to draw conclusions about a population. Specifically, since we have two populations that need to be compared, we leverage on *one-sided paired samples t-test*. The paired t-test determines whether the mean change

TABLE 2. Confidence intervals of the Dice test metric of PosPadUNet3D and variations of our MATAP, both trained on the ToothFairy dataset for the generation phase only.

Method	Transf.	ABS Token	Memory	Hann Window	Lower Bound	Mean	Upper Bound	STD
PosPadUNet3D	✗	✗	0	✗	0.792	0.797	0.801	0.006
TransPosPadUNet3D	✓	✗	0	✗	0.790	0.796	0.802	0.009
TransPosPadUNet3D	✓	✓	0	✗	0.797	0.801	0.804	0.005
TransPosPadUNet3D	✓	✗	128	✗	0.792	0.800	0.807	0.011
TransPosPadUNet3D	✓	✓	128	✗	0.799	0.802	0.805	0.004
MATAP	✓	✓	128	✓	**0.806**	**0.809**	**0.812**	0.004

TABLE 3. One-sided paired samples t-test of the Dice test metric of PosPadUNet3D, variants of MATAP trained on the ToothFairy dataset for the generation phase only.

First Population	Second Population	p-value
PPUNet	TPPUNet + ABS	6.3×10^{-2}
PPUNet	TPPUNet + ABS + Mem.	1.2×10^{-2}
PPUNet	MATAP	2.8×10^{-5}

between the two populations under examination is significantly different from zero. Hence, it can be used to determine if one mean is consistently greater than the other. By means of this test, we can statistically demonstrate the superiority of our proposed approach compared to the existing literature.

Formally, we define our hypothesis test as follows:

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \\ H_A &: \mu_X < \mu_Y \end{aligned} \tag{8}$$

where H_0 states that the mean of the first $(X_1, \ldots, X_m)$ and second $(Y_1, \ldots, Y_n)$ population are equal, and our objective is to disprove such a hypothesis in favor of the alternative, stating that the mean of the first population is significantly smaller than that of the second one.

In our context, we can mathematically define the degrees of freedom of the t-test distribution, ν, and the test statistic value, t, as follows:

$$\nu = \left\lfloor \frac{(\frac{S_X^2}{m} + \frac{S_Y^2}{n})^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}} \right\rfloor, \quad t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \tag{9}$$

where S is the *unbiased estimator* of the standard deviation introduced in Eq. (4).

Finally, we compute the area under the ν-degrees of freedom t-curve to the right of t value find with Eq. (9). Such area represents the so-called *p-value* and corresponds to the probability of obtaining a t that is greater, or at least equal, to the one that is actually observed, assuming that the null hypothesis is true. In other words, a smaller p-value means that there is stronger evidence in favor of the alternative hypothesis, in our case the distribution $\mu_X < \mu_Y$. The detailed outcomes of these hypothesis tests are presented in Tab. 3.

V. EXPERIMENTS & RESULTS

According to our evaluation protocol described in Sec. IV-E, with the aim of ensuring the robustness of the proposed model and obtaining reliable performance estimates, we repeated the training procedure 10 times using different random seeds. By doing so, we are able to compute confidence intervals and hypothesis testing, hence obtaining a comprehensive assessment of MATAP's performance. Hyperparameters are fixed among all the different experiments: we set the batch size to 2, used Adam as optimizer with a learning rate 0.0001, a weight decay of 0.00005, and no momentum. We employed the soft IoU as loss and trained each phase for 100 epochs. The size of the extracted patches was (120, 120, 120) for the Generation phase, and (80, 80, 80) for the other phases. For every other detail about the experiment configuration, we suggest taking a look at the configuration files in our public repository, which briefly describe every minimal tuning adopted.[2]

A. ON THE EFFECTIVENESS OF ABS TOKEN AND MEMORY

To show the contribution of each model component, we start our evaluation by progressively including them in Tab. 2. In order to keep the number of training procedures reasonably modest, we performed our experiments focusing only on the generation phase of the training (Sec. IV-D). It is worth noticing that any improvement in this step will benefit the whole segmentation pipeline. In the aforementioned table, the confidence intervals of the test Dice metric are reported for each experiment. Additionally, Tab. 3 reports the paired samples t-tests to assess whether our proposal is statistically valuable or not. Such a table should be read as follows: a smaller p-value means that there is stronger evidence in favor of the alternative hypothesis, meaning in our case that the mean of the first population is smaller than the mean of the second. Our populations are composed of the Dice metrics computed on the test set with the predictions of a model trained 10 times under the same experimental conditions, except only for random seeds.

As mentioned, Tab. 2 evaluates the contribution of each component to our proposal by increasingly introducing it onto the baseline, PosPadUNet3D [35]. At first glance, the comparison between the first two lines of the table might imply a lack of efficacy of the Transformer architecture. However, it is crucial to note that PosPadUNet3D incorporates absolute positional information from the original volume, which is not the case for TransPosPadUNet3D, which simply relies on a Transformer module introduced in the bottleneck of the U-Net architecture.

[2]https://github.com/AImagelab-zip/alveolar_canal

TABLE 4. State-of-the-art comparison on the Maxillo dataset. Missing results were not provided by the original papers.

Method	IoU	Dice
Liu et al. [40]	–	0.756
Usman et al. [38]	–	0.770
Cripriano et al. [35]	0.650	0.790
Zhao et al. [39]	–	0.810
Ours	**0.704**	**0.824**
Ours (training on ToothFairy)	**0.710**	**0.831**

Introducing the `[ABS]` token to TransPosPadUNet3D (third line of Tab. 2) enhances its performance, already improving with respect to PosPadUNet3D, as statistically evidenced also in Tab. 3. Furthermore, the performance of TransPosPadUNet3D shows a progressive improvement, initially with the integration of memory tokens, and subsequently through the application of the Hann Windows function as a post-processing strategy. Ultimately, the implementation of the `[ABS]` token results in a halved standard deviation, thereby supporting the enhanced robustness of the proposed model.

B. ABOUT COMPUTATIONAL REQUIREMENTS

Regarding the convergence speed and computational time, it is worth recalling that all presented architecture variations are based on the U-Net framework. These architectures slightly differ in the number of layers and hyperparameters without any significant influence on the convergence speed and computational time. The only exception is MATAP, which requires additional parameters due to the integration of a transformer in the bottleneck. This modification has resulted in a measurable increase in the training computational time of approximately 21%. The most notable difference regards the number of parameters, that is double with respect to the base U-Net model (55.24 million parameter instead of the 20.19 million of PosPadUNet3D).

Nevertheless, the inference time is approximately the same, also when considering the MATAP model. As an example, on a NVIDIA GeForce RTX 2080 Ti with 12GB of VRAM, all the architecture variations analyzed in the paper require less than 3s for processing a $170 \times 340 \times 370$ input volume, making the tools suitable for a clinical application and significantly reducing the segmentation time required in the daily practice. Indeed, if a manual 2D sparse segmentation requires about 2 minutes to be completed, with the proposed tool it is possible to obtain a 3D dense segmentation in a matter of a few seconds.

C. COMPARISON WITH THE STATE OF THE ART

In order to compare our proposal with the latest advances on the segmentation of the inferior alveolar nerve [38], [39], [40], Tab. 4 is also provided. Both [38] and [39] leverage a two-stage approach that aims at filtering out background data before actually performing the canal segmentation. In doing so, [38] makes use of a CNN-based approach that performs worse than both the positional encoding proposed in [35], and the non-deep two-stage approach based on the Frenet frame described in [39]. Liu et al. [40], the worst among the considered competitors, is again based on a U-Net architecture enhanced with a frequency attention module.

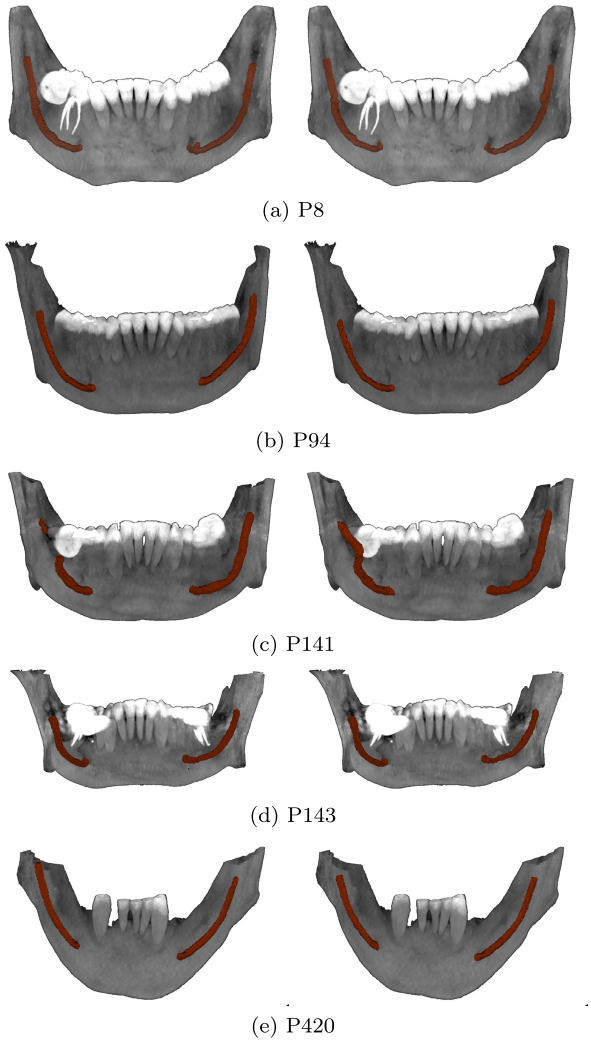


FIGURE 8. Pair of predictions made by our proposed model MATAP (left) and ground-truths on examples taken from the test set (right). The jaws face the camera view, thus the canal on the left side is the right IAC.

To make the evaluation as fair as possible, the comparison reported in Tab. 4 is based on the Maxillo dataset only, which was the target dataset for both [38] and [39]. The employed training procedure was described in Section IV-D. Such a comparative evaluation confirms that our proposed model, MATAP, outperforms the state-of-the-art competitors on the public dataset.

To provide the reader with a complete evaluation, we also performed the complete training procedure on the ToothFairy dataset, obtaining an overall segmentation score of 0.831 Dice and 0.710 IoU.

D. QUALITATIVE EVALUATION

To provide a visual representation of the predictions obtained using our MATAP model, Fig.8 showcases five pairs of

automatic segmentations coupled with their corresponding ground-truth annotations. Sample patients are taken from the public test case of the ToothFairy dataset.

While the majority of the predictions are exceptionally accurate and worth to be integrated in the daily clinical practice, a notable edge case is observed in the sample P141, where the canal on the left is heavily affected by the presence of a wisdom tooth, making it one of the hardest to be predicted. In this instance, our model's prediction resulted in a non-continuous canal. Further improvements of our MATAP may involve techniques to deal with such a kind of issues.

VI. CONCLUSION

In conclusion, this paper introduces a novel approach for segmenting the inferior alveolar canal. Our approach addresses the limitations related to patch-based learning by incorporating the global coordinates of each extracted patch into a transformer-based architecture. The proposed design choice enhances the model's ability to efficiently leverage global spatial information by projecting patch coordinates into the input sequence of the transformer architecture. Additionally, we introduce post-processing techniques based on the Hann window function to effectively remove artifacts that arise at patch borders. The achieved state-of-the-art results are consistently demonstrated across multiple experimental runs and their statistical significance is validated using the Student's t-distribution.

To ensure the reproducibility of our experiments, we have made the described pipelines openly accessible to the scientific community as an open-source project. Furthermore, we conducted our experiments on public datasets encouraging the scientific community to further enhance the results in the context of inferior alveolar canal segmentation and letting anyone reproduce the obtained results and verify our claims. Such a collaborative effort is crucial in medical-related critical domains to foster progress and innovation.

While the suggested approach has proven effective in refining IAC segmentation, it could be adapted and potentially applied to any tasks where feeding an entire sample into the network is impractical, but having a global context is important.

Future works will focus on studying the versatility of our proposed method which would open doors to a broader range of applications beyond IAC segmentation, offering a promising research direction for further investigation into its performance across diverse neural networks, datasets, and data modalities.

REFERENCES

[1] G. Juodzbalys, H.-L. Wang, and G. Sabalys, "Injury of the inferior alveolar nerve during implant placement: A literature review," *J. Oral Maxillofacial Res.*, vol. 2, no. 1, pp. 1–12, Jan. 2011.

[2] A. Westermark, S. Zachow, and B. L. Eppley, "Three-dimensional osteotomy planning in maxillofacial surgery including soft tissue prediction," *J. Craniofacial Surg.*, vol. 16, no. 1, pp. 100–104, Jan. 2005.

[3] M. D. Bartolomeo, A. Pellacani, F. Bolelli, M. Cipriano, L. Lumetti, S. Negrello, S. Allegretti, P. Minafra, F. Pollastri, R. Nocini, G. Colletti, L. Chiarini, C. Grana, and A. Anesi, "Inferior alveolar canal automatic detection with deep learning CNNs on CBCTs: Development of a novel model and release of open-source dataset and algorithm," *Appl. Sci.*, vol. 13, no. 5, p. 3271, 2023.

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–12.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[9] F. Pollastri, M. Parreño, J. Maroñas, F. Bolelli, R. Paredes, D. Ramos, and C. Grana, "A deep analysis on high-resolution dermoscopic image classification," *IET Comput. Vis.*, vol. 15, no. 7, pp. 514–526, Oct. 2021.

[10] S. Soffer, A. Ben-Cohen, O. Shimon, M. M. Amitai, H. Greenspan, and E. Klang, "Convolutional neural networks for radiologic images: A radiologist's guide," *Radiology*, vol. 290, no. 3, pp. 617–627, 2019.

[11] F. Pollastri, J. Maroñas, F. Bolelli, G. Ligabue, R. Paredes, R. Magistroni, and C. Grana, "Confidence calibration for deep renal biopsy immunofluorescence image classification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1298–1305.

[12] M. Lovino, M. Montemurro, V. S. Barrese, and E. Ficarra, "Identifying the oncogenic potential of gene fusions exploiting miRNAs," *J. Biomed. Informat.*, vol. 129, May 2022, Art. no. 104057.

[13] P. Barbiero, M. Lovino, M. Siviero, G. Ciravegna, V. Randazzo, E. Ficarra, and G. Cirrincione, "Unsupervised multi-omic data fusion: The neural graph learning network," in *Proc. Int. Conf. Intell. Comput.*, 2020, pp. 172–178.

[14] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of Swin transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20698–20708.

[15] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1748–1758.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 1–6.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] A. Schramm, M. Rucker, N. Sakkas, R. Schon, J. Duker, and N.-C. Gellrich, "The use of cone beam CT in cranio-maxillofacial surgery," in *Proc. Int. Congr. Ser.*, vol. 1281, 2005, pp. 1200–1204.

[19] T. Kondo, S. H. Ong, and K. W. C. Foong, "Computer-based extraction of the inferior alveolar nerve canal in 3-D space," *Comput. Methods Programs Biomed.*, vol. 76, no. 3, pp. 181–191, Dec. 2004.

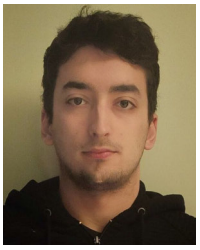
[20] S. Rueda, J. A. Gil, R. Pichery, and M. Alcañiz, "Automatic segmentation of jaw tissues in CT using active appearance models and semi-automatic landmarking," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2006, pp. 167–174.

[21] W. C. Scarfe, A. G. Farman, and P. Sukovic, "Clinical applications of cone-beam computed tomography in dental practice," *J. Can. Dental Assoc.*, vol. 72, pp. 75–80, Sep. 2006.

[22] D. Kainmueller, H. Lamecker, H. Seim, M. Zinser, and S. Zachow, "Automatic extraction of mandibular nerve and bone from cone-beam CT data," in *Proc. Int. Conf. Med. Image Comput., Comput.-Assist. Intervent*, 2009, pp. 76–83.

[23] D.-J. Kroon, "Segmentation of the mandibular canal in cone-beam CT data," Ph.D. dissertation, Univ. Twente, Enschede, Netherlands, 2011, doi: 10.3990/1.9789036532808.

[24] F. Abdolali and R. A. Zoroofi, "Mandibular canal segmentation using 3D active appearance models and shape context registration," in *Proc. 21st Iranian Conf. Biomed. Eng. (ICBME)*, Nov. 2014, pp. 7–11.
[25] F. Abdolali, R. A. Zoroofi, M. Abdolali, F. Yokota, Y. Otake, and Y. Sato, "Automatic segmentation of mandibular canal in cone beam CT images using conditional statistical shape model and fast marching," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 12, no. 4, pp. 581–593, Apr. 2017.
[26] B. Moris, L. Claesen, Y. Sun, and C. Politis, "Automated tracking of the mandibular canal in CBCT images using matching and multiple hypotheses methods," in *Proc. 4th Int. Conf. Commun. Electron. (ICCE)*, Aug. 2012, pp. 327–332.
[27] X. Wei and Y. Wang, "Inferior alveolar canal segmentation based on cone-beam computed tomography," *Med. Phys.*, vol. 48, no. 11, pp. 7074–7088, Nov. 2021.
[28] J. Blacher, S. Van DaHuvel, V. Parashar, and J. C. Mitchell, "Variation in location of the mandibular foramen/inferior alveolar nerve complex given anatomic landmarks using cone-beam computed tomographic scans," *J. Endodontics*, vol. 42, no. 3, pp. 393–396, Mar. 2016.
[29] J.-J. Hwang, Y.-H. Jung, B.-H. Cho, and M.-S. Heo, "An overview of deep learning in the field of dentistry," *Imag. Sci. Dentistry*, vol. 49, no. 1, p. 1, 2019.
[30] G. H. Kwak, E.-J. Kwak, J. M. Song, H. R. Park, Y.-H. Jung, B.-H. Cho, P. Hui, and J. J. Hwang, "Automatic mandibular canal detection using a deep convolutional neural network," *Sci. Rep.*, vol. 10, no. 1, p. 5711, Mar. 2020.
[31] J. Jaskari, J. Sahlsten, J. Järnstedt, H. Mehtonen, K. Karhu, O. Sundqvist, A. Hietanen, V. Varjonen, V. Mattila, and K. Kaski, "Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes," *Sci. Rep.*, vol. 10, no. 1, pp. 1–8, Apr. 2020.
[32] J. Järnstedt, J. Sahlsten, J. Jaskari, K. Kaski, H. Mehtonen, A. Hietanen, O. Sundqvist, V. Varjonen, V. Mattila, S. Prapayasatok, and S. Nalampang, "Reproducibility analysis of automated deep learning based localisation of mandibular canals on a temporal CBCT dataset," *Sci. Rep.*, vol. 13, no. 1, p. 14159, Aug. 2023.
[33] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[34] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 424–432.
[35] M. Cipriano, S. Allegretti, F. Bolelli, F. Pollastri, and C. Grana, "Improving segmentation of the inferior alveolar nerve through deep label propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21105–21114.
[36] P. Lahoud, S. Diels, L. Niclaes, S. Van Aelst, H. Willems, A. Van Gerven, M. Quirynen, and R. Jacobs, "Development and validation of a novel artificial intelligence driven tool for accurate mandibular canal segmentation on CBCT," *J. Dentistry*, vol. 116, Jan. 2022, Art. no. 103891.
[37] C. Mercadante, M. Cipriano, F. Bolelli, F. Pollastri, M. D. Bartolomeo, A. Anesi, and C. Grana, "A cone beam computed tomography annotation tool for automatic detection of the inferior alveolar nerve canal," in *Proc. 16th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2021, pp. 724–731.
[38] M. Usman, A. Rehman, A. M. Saleem, R. Jawaid, S.-S. Byon, S.-H. Kim, B.-D. Lee, M.-S. Heo, and Y.-G. Shin, "Dual-stage deeply supervised attention-based convolutional neural networks for mandibular canal segmentation in CBCT scans," *Sensors*, vol. 22, no. 24, p. 9877, Dec. 2022.
[39] H. Zhao, J. Chen, Z. Yun, Q. Feng, L. Zhong, and W. Yang, "Whole mandibular canal segmentation using transformed dental CBCT volume in frenet frame," *Heliyon*, vol. 9, no. 7, Jul. 2023, Art. no. e17651.
[40] Z. Liu, D. Yang, M. Zhang, G. Liu, Q. Zhang, and X. Li, "Inferior alveolar nerve canal segmentation on CBCT using U-Net with frequency attentions," *Bioengineering*, vol. 11, no. 4, p. 354, Apr. 2024.
[41] M. Cipriano, S. Allegretti, F. Bolelli, M. D. Bartolomeo, F. Pollastri, A. Pellacani, P. Minafra, A. Anesi, and C. Grana, "Deep segmentation of the mandibular canal: A new 3D annotated dataset of CBCT volumes," *IEEE Access*, vol. 10, pp. 11500–11510, 2022.
[42] L. Lumetti, V. Pipoli, F. Bolelli, and C. Grana, "Annotating the inferior alveolar canal: The ultimate tool," in *Proc. Int. Conf. Image Anal. Process.*, 2023, pp. 525–536.
[43] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Trans. Inf. Theory*, vols. IT-29, no. 4, pp. 551–559, Jul. 1983.
[44] G. Bontempo, A. Porrello, F. Bolelli, S. Calderara, and E. Ficarra, "DAS-MIL: Distilling across scales for MIL classification of histological WSIs," in *Proc. Int. Conf. Med. Image Comput., Comput.-Assist. Intervent*, 2023, pp. 248–258.
[45] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10575–10584.
[46] N. Pielawski and C. Wählby, "Introducing Hann windows for reducing edge-effects in patch-based image segmentation," *PLoS One*, vol. 15, no. 3, Mar. 2020, Art. no. e0229839.



LUCA LUMETTI received the B.Sc. and M.Sc. degrees in computer engineering from the Università degli Studi di Modena e Reggio Emilia, Italy, where he is currently pursuing the Ph.D. degree with the AImageLab Group. His research interests include artificial intelligence, computer vision, and medical imaging.

VITTORIO PIPOLI received the B.Sc. degree in computer engineering and the M.Sc. degree in data science and engineering from the Politecnico di Torino, Italy. He is currently pursuing the Ph.D. degree with the AImageLab Group, Università degli Studi di Modena e Reggio Emilia. His research interests include artificial intelligence, computer vision, and medical imaging.

FEDERICO BOLELLI (Associate Member, IEEE) received the B.Sc. and M.Sc. degrees in computer engineering and the Ph.D. degree from Università degli Studi di Modena e Reggio Emilia, Italy. He is currently an Assistant Professor with the AImageLab Group, Dipartimento di Ingegneria "Enzo Ferrari," Università degli Studi di Modena e Reggio Emilia. He is also involved in a H2020 European projects. His research interests include image processing, algorithms and optimization, and medical imaging.

ELISA FICARRA received the Ph.D. degree in systems and computer engineering from the Politecnico di Torino, Turin, Italy, in 2006. She is currently a Full Professor with the Dipartimento di Ingegneria "Enzo Ferrari," Università degli Studi di Modena e Reggio Emilia, Italy. Her research interests include biological image processing and bioinformatics, high-throughput sequencing analysis, and artificial intelligence for biological and smart manufacturing applications.

COSTANTINO GRANA received the degree from Università degli Studi di Modena e Reggio Emilia, Italy, in 2000, and the Ph.D. degree in computer science and engineering, in 2004. He is currently a Full Professor with the Dipartimento di Ingegneria, Università degli Studi di Modena e Reggio Emilia. He has published six book chapters, 47 articles on international peer-reviewed journals, and more than 130 papers on international conferences. His research interests include medical imaging, optimization of image processing algorithms, and computer vision applications.

Open Access funding provided by 'Università degli Studi di Modena e Reggio Emilia' within the CRUI CARE Agreement