Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs

Luca Lumetti[1], Vittorio Pipoli[1,2], Federico Bolelli[1], Elisa Ficarra[1], and Costantino Grana[1]

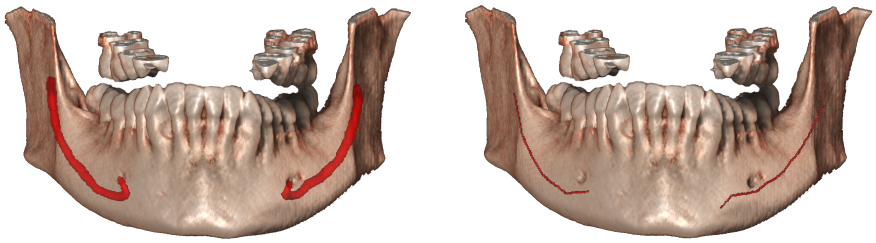
[1] University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it
[2] University of Pisa, Italy
{name.surname}@phd.unipi.it

Abstract. The segmentation of the Inferior Alveolar Canal (IAC) plays a central role in maxillofacial surgery, drawing significant attention in the current research. Because of their outstanding results, deep learning methods are widely adopted in the segmentation of 3D medical volumes, including the IAC in Cone Beam Computed Tomography (CBCT) data. One of the main challenges when segmenting large volumes, including those obtained through CBCT scans, arises from the use of patch-based techniques, mandatory to fit memory constraints. Such training approaches compromise neural network performance due to a reduction in the global contextual information. Performance degradation is prominently evident when the target objects are small with respect to the background, as it happens with the inferior alveolar nerve that develops across the mandible, but involves only a few voxels of the entire scan.
In order to target this issue and push state-of-the-art performance in the segmentation of the IAC, we propose an innovative approach that exploits spatial information of extracted patches and integrates it into a Transformer architecture. By incorporating prior knowledge about patch location, our model improves state of the art by ∼2 points on the Dice score when integrated with the standard U-Net architecture. The source code of our proposal is publicly released.

Keywords: Inferior Alveolar Canal · 3D Segmentation · CBCT · Transformers · Patch-based Learning

1 Introduction

The presence of the Inferior Alveolar Nerve (IAN) represents a challenge for maxillofacial surgery. Such a nerve crosses the Inferior Alveolar bone Canal (IAC) and supplies sensation to the lower teeth, lips, and chin. For this reason, IAN position (Fig. 1) must be carefully identified before surgical intervention (e.g., implant placement and molar extraction) to prevent aches, pain, and temporary or permanent paralysis [33]. Usually, the preoperative treatment planning



(a) (b)

Fig. 1. CBCT with the IAC marked in red. (a) contains a 3D dense annotation, while (b) contains a 2D sparse annotation obtained from a panoramic view of the mandible and later re-projected to the 3D space.

is based on IAC segmentation performed on 3D data acquired with Cone Beam Computer Tomography (CBCT). Nevertheless, producing 3D annotations for 3D data is dramatically challenging and time-consuming. Hence, the standard practice consists of extracting 2D panoramic views where the surgeon can annotate the approximate position of the IAC drawing 2D curves. Despite this procedure being effective most of the time, having the 3D segmentation of the IAC would crucially improve the precision of the surgery planning, minimizing the likelihood of errors during surgery operations.

Recent advancements in deep learning have significantly impacted multiple domains, including medical imaging, particularly through methods based on Convolutional Neural Networks (CNNs) [11, 14, 15, 23–26]. Among them, of the most popular is U-Net [27], an encoder-decoder architecture with skip connections capable of extracting deep features while trying to retain as many fine-grained details as possible [10]. As well, many U-Net-based approaches for the automatic segmentation of the IAC [5, 16, 30] have been recently published, also thanks to the public availability of a 3D-annotated dataset [4].

Despite the great success of CNN in medical imaging, the rise of Transformer architectures [31] stands as a turning point. Representing the standard of Natural Language Processing since 2017 and deeply affecting the Computer Vision field since 2020 [8], Transformer-based architectures demonstrate dominance in several tasks due to their capability of modeling long-range interactions [6,21,22,28]. This is in contrast with the *CNN locality bias*, which instead forces the modeling of local interactions that lie within the CNN sliding kernels [8]. For this reason, researchers are developing strategies to improve U-Net-based architectures [27] by integrating some Transformer layers to enhance long-range interactions, with encouraging results [9, 29]. In this work, we investigate innovative and effective ways to improve such an integration.

Regardless of the adotped method, processing 3D volumes leads to severe memory constraints, making the segmentation of a single 3D scan in one shot a prohibitive operation. Meanwhile, decreasing the resolution of such 3D images

with downsampling techniques is counterproductive because fine-grained details are needed to improve the segmentation quality. Hence, the only solution to solve both problems is splitting the 3D scans into multiple patches that will be processed separately, without losing detailed information. The literature refers to the aforementioned procedure as *patch-based learning*. Even if patch-based learning allows the training of deep neural networks with standard hardware resources, it must be mentioned that it forces the model to focus only on a fraction of the total information at a time, losing global context (e.g., the position of the examined patch with respect to the other patches of the 3D volume). In this research, we aim at mitigating this phenomenon with Transformers [31].

Paper Contribution. We present an innovative 3D segmentation model enhanced by a memory-augmented Transformer encoder that effectively harnesses absolute spatial coordinates, addressing the challenges of patch-based training.

Specifically, our proposal evolves from the standard 3D U-Net architecture by incorporating a memory-augmented Transformer in the bottleneck. By leveraging 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, thereby increasing contextualization. Moreover, we harness such a flow of information to effectively inject contextual information related to the processed patches, i.e., their position within the entire volume, thus mitigating issues associated with patch-based learning. The "memory" is an additional refinement that supports the model in retaining crucial prior concepts that may be challenging to be directly extracted from image features, but are nonetheless valuable for interpretation. In summary, the key contributions of this paper are as follows:

i) We propose a memory-augmented Transformer module that harnesses absolute spatial coordinates, mitigating issues related to patch-based learning;
ii) We design an U-Net-based deep learning architecture integrating our proposed module and tailored for 3D IAC segmentation, outperforming state of the art on the selected segmentation task of ∼2 Dice points;
iii) The source code of our proposal is publicly released[3] to allow the replication of the experiments and foster future research advancements.

2 Related Works

While classical computer vision approaches have made significant contributions in the past [1,2,13,19,32], today, the most successful models for the segmentation of the IAC are based on machine learning and deep learning.

Notably, Jaskari *et al.* [12] presented one of the pioneering applications of deep learning for mandibular canal segmentation. Their approach involved training a convolutional network using a dataset of coarsely annotated 3D scans. This

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

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.

Cipriano *et al.* [5] introduced a significant breakthrough by proposing the first publicly available dataset of 3D annotated CBCT scans of the human jaw, named *Maxillo*, alongside a deep learning model for the 3D segmentation of the IAC, *PosPadUNet3D*. This marks a substantial advancement in publicly accessible datasets for the segmentation of the inferior alveolar canal. The Maxillo dataset has been later extended with the 2023 MICCAI ToothFairy Challenge.[4]

Additionally, in [30], Usman *et al.* proposed a two-stage approach also based on the U-Net architecture. 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, they initially apply a CNN to isolate the regions of the input volume where the canal is likely to be located, reducing background interference. Then, leveraging U-Net architecture, the segmentation of the mandibular canal is performed exclusively within the extracted regions.

The latest approach tested on public data is contributed by Zhao *et al.* [34] and, similarly to [30], it works in a two-stage fashion. Firstly, the mandibular centerline is extracted via automatic segmentation of the mandible and localization of the mandibular and mental foramen. The sub-volumes containing the mandibular canal information are then obtained using a double reflection method based on the Frenet frame. Secondly, the extracted sub-volumes are fed into a U-Net-based 3D segmentation network, and the topology of the mandibular canal is constrained with the clDice. To conclude the segmentation process, the prediction masks are inversely transformed back into the original CBCT images.

2.1 Patch-based Learning

All of the aforementioned solutions employ a patch-based learning strategy. Indeed, when targeting complex, high-dimensional inputs or when the computation resources available are limited or should be kept so, patch-based learning is the only viable approach. 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 meet memory constraints, the simple downsampling of the input data is counterproductive whenever the preservation of fine-grained details is crucial. A common approach consists of training neural networks using subsets extracted from the original data [3, 12]. Such an approach, known as patch-based learning, mitigates memory constraints but also leads to a loss of global information due to restricted (patch-limited) receptive fields. Moreover, ambiguity in segmenting

[4] https://toothfairy.grand-challenge.org

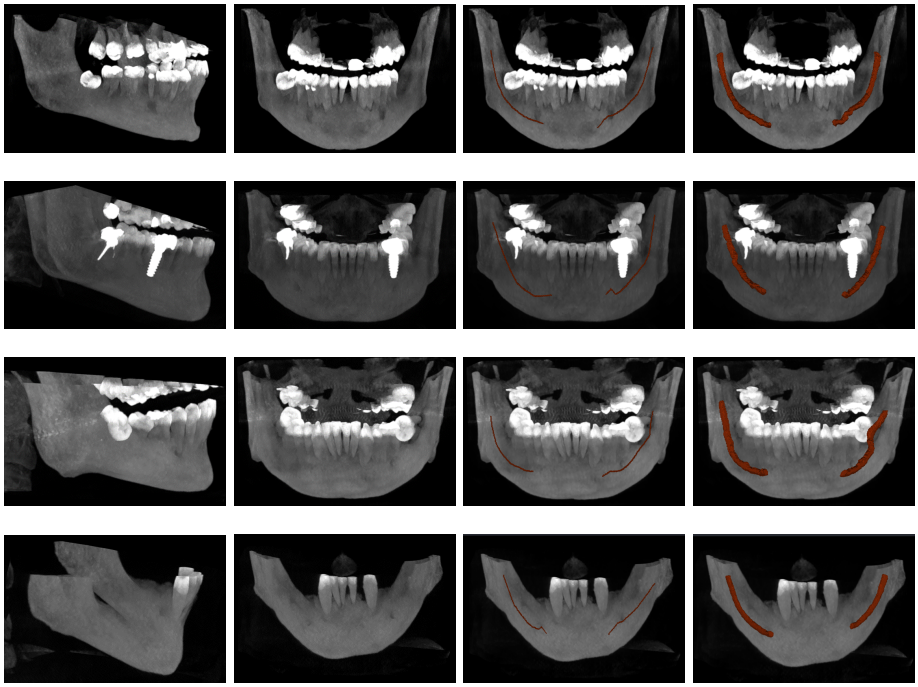


Fig. 2. Sample data from the ToothFairy dataset. Each line of the image contain a different patient, from left to right you can see left-side and frontal views of the CBCT volume, sparse and dense annotations of the inferior alveolar nerve.

objects situated at the intersections of multiple patches may arise, causing potential artifacts around patch boundaries. When the object to be segmented is small in comparison to the entire volume, as it happens in the segmentation of the IAC, the aforementioned challenges become particularly prominent.

A first proposal to overcome the patch-based learning drawbacks in the segmentation of the IAN is introduced in [5] 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 performance, the aforementioned major issues still persisted. Unlike *PosPadUNet3D*, our approach harnesses the information flow of Transformers, semantically conditioning the bottleneck embedding based on the spatial information instead of a simple feature concatenation.

3 Dataset

The maxillofacial dataset employed in our experiments is an improved version of the *Maxillo* dataset introduced by Cipriano *et al.* [4]. Such an improvement,

known as *ToothFairy* dataset, was part of the homonymous MICCAI 2023 challenge hosted on the *Grand Challenge* platform.[5]

All of the 3D CBCT volumes of the ToothFairy dataset were collected from the Affidea center in Modena, Italy, part of a leading pan-European healthcare group specializing in advanced diagnostics, outpatient services, laboratory analyses, physiotherapy and rehabilitation, and cancer diagnosis and treatment. The scans were acquired using the NewTom/NTVGiMK4 CBCT device, with acquisition parameters set at 3 mA, 110 kV, and 0.3 mm cubic voxels. The dataset is publicly available after user registration:[6] such availability, along with the public release of the source code, ensures full reproducibility of our experiments and verification of our claims.

The annotation process was initially performed by diagnostic technicians responsible for the examinations, providing what we refer to as "sparse annotations" (Fig. 1b): the upper boundary of the canal is marked along the entire dental arch, offering a useful approximation of the nerve position. Such annotations 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.

Instead, the 3D annotations (in the following also referred to as "dense annotations") of the ToothFairy dataset (Fig. 1a) have been created using an updated version of the IACAT tool [18], specifically version 2.0 developed in [17], by a team of medical experts with over five years of experience in the maxillofacial field.

All of the 443 volumes in the ToothFairy dataset are paired with the 2D sparse annotation. For a subset of 153 scans, the 3D dense annotation is also provided. For what concerns volume shapes, the average size in the dataset is $169 \times 342 \times 370$, while minimum and maximum volumes have respectively $148 \times 265 \times 312$ and $178 \times 423 \times 463$ dimensions. Sample images of the employed dataset are reported in Fig. 2.

4 Methods

This paper proposes a novel U-Net-based deep learning model for the segmentation of the IAC. Specifically, we devise a module that harnesses memory-augmented Transformer layers for modeling long-range interactions and integrating absolute positional information to mitigate issues related to patch-based learning. All the details concerning our proposed methodology can be found in Sec. 4.1.

In our work, a two-step training procedure is employed to exploit both volumes that are annotated in 3D and those that are annotated only in 2D, improving overall segmentation performances. An in-depth explanation of this training procedure can be found in Sec. 4.2.

Finally, the Hann-based post-processing employed in our pipeline is described in Sec. 4.3.

[5] https://toothfairy.grand-challenge.org/
[6] https://ditto.ing.unimore.it/toothfairy/

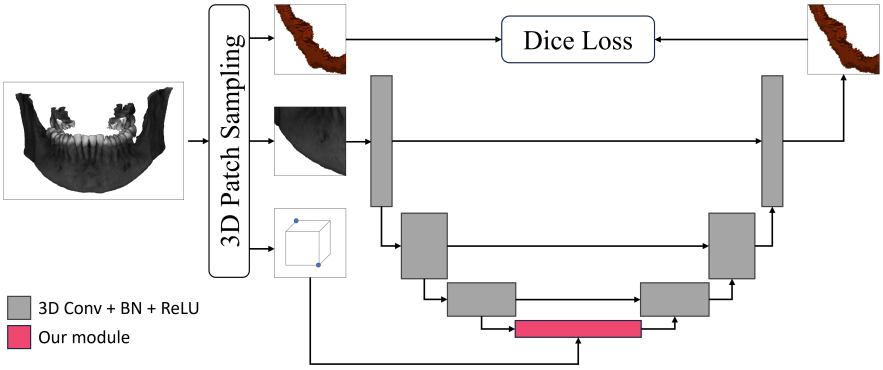


Fig. 3. Proposed Transformer module integrated in the standard 3D U-Net architecture. A detailed visualization of our module is reported in Fig. 4.

4.1 The Proposed Approach

We design a novel deep-learning model to address the limitations associated with patch-based learning through the utilization of Transformers capable of exploiting contextual information. More specifically, we propose a module based on Transformer encoder blocks, accompanied by learned embedding representations for positional encoding, and integrate it in the bottleneck of the well-known U-Net architecture (Fig. 3). 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 bottleneck dimensional space, exploiting a learnable matrix of dimension $6 \times d_{model}$, where 6 are the numbers identifying the position of the patch within the entire volume and d_{model} is the number of channels in the U-Net bottleneck (Fig. 4). 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 of the other tokens should be able to employ its information independently from their position. To achieve this goal, the Transformer inbuilt positional encoding is summed only to the tokens representing volume information. Again, this ensures that the `[ABS]` token remains positionally untied from the rest of the sequence. This

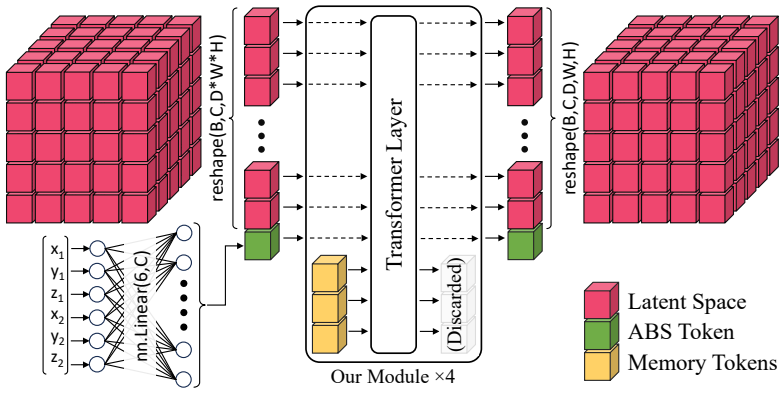


Fig. 4. The proposed module. `B`, `C`, `D`, `W`, and `H` represent respectively batch size, channels, depth, height, and width. The patch coordinates $[x_1, y_1, z_1, x_2, y_2, z_2]$ are projected using a linear layer to produce the `[ABS]` token. The activation map obtained in the bottleneck of U-Net before the first transposed convolution (pink blocks) is flattened across the spatial dimension and concatenated with the `[ABS]` token. The resulting tensor is fed to a cascade of four Transformer layers: for each layer a new set of memory token is concatenated to the input sequence and discarded from the output so that the sequence length does not vary. After the Transformer layers, the `[ABS]` token is removed and the remaining output is reshaped back to the original spatial dimensionality.

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, we enriched the proposed module with *memory*. The integration of Transformer memory has demonstrated considerable effectiveness in tasks such as image captioning [7]. This mechanism enables the Transformer to retain crucial prior 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. A graphical summary of this process is provided in Fig. 4.

4.2 Model Training

With the aim of also leveraging volumes with only 2D sparse annotations available, we adopt a 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 dense 3D annotations. Again, the rationale behind this operation is to obtain a model that can leverage the sparse 2D labels (available for all the volumes in

the Maxillo dataset) 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 with the synthetically annotated CBCT volumes, generated by the deep label expansion. Thus, a total of 420 3D annotated volumes is obtained as a train set. Still, 8 and 15 volumes from the non-synthetic 3D dataset are available for the validation and test set respectively. Using the above-mentioned set of data, our segmentation model is trained to output 3D masks representing the inferior alveolar canal, and consequentially evaluated.

In other words, our pipeline leverage the proposed model twice, changing only the input data. A first instance is used to extend the amount of 3D IAC annotations by learning to "expand" the available 2D labels. The second instance is trained to predict a 3D segmentation of the IAC starting from a virgin scan. Test data of both instances are never seen during training.

4.3 Post-Processing

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 on patch borders, we still need to deal with noise and artifacts generated at patch boundaries (Fig. 5). Taking inspiration from audio encoding [20], 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:

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

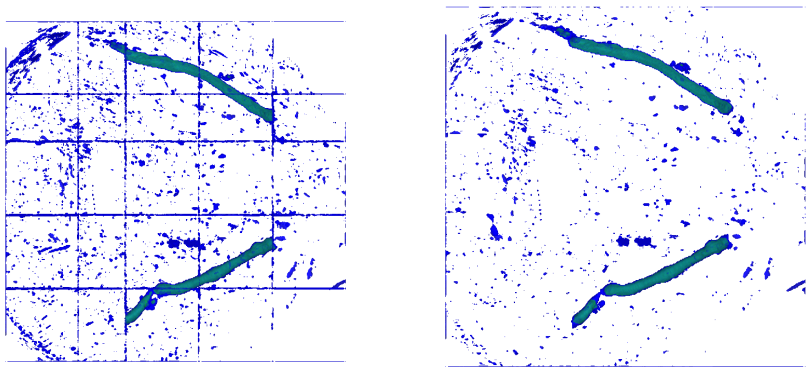
where i is an element in the considered interval I. This function is symmetric, peaking at 1 in the middle of the window and tapering to 0 at the edges.The sum of two Hann windows, each shifted by $\frac{I}{2}$ (50%), is equivalent to a rectangular window of width I and height 1:

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

Such a property is exploited in audio encoding to eliminate border artifacts by multiplying the Hann window with frames that overlap by 50%, before summing them together.

While this approach is defined in 1D for audio, we extended it to multiple dimensions, and applied it to the 3D segmented patches produced by our model. Thus, we are able to reduce the aforementioned noise on patch borders and improve the overall performance.

The effects of the proposed 3D extension of the Hann filtering to the segmented patches produced by our model are depicted in Fig. 5.



(a) W/O Post-Processing (b) Hann-based Filtering

Fig. 5. Effects of Hann-based filtering on an axial plane extracted from a predicted volume. On the left (a), the prediction of our model without post-processing. On the right (b), the effect of the proposed Hann-based post-processing on the same model output. In both images, blue represents logits that have a value higher than 10^{-4}. The post-processing significantly reduces artifacts that appear close to patch borders. Even if most of these artifacts do not cause any issues, the ones that are close to the IAC badly influence the final segmentation.

5 Experiments and Results

Sec. 5.1 defines the details of the adopted patch-based learning procedure, alongside with our experimental setting. We compare our proposal with state-of-the-art models in Sec. 5.2 and conduct an ablation study to highlight the contribution of the absolute token `[ABS]` and the memory of the Transformer in Sec. 5.3. Finally, Sec. 5.4 provides some visualizations of our model predictions, discussing its strengths and weaknesses.

5.1 Experimental Setting

Since we adopted a patch-based learning approach, we fed our model with patches of $120 \times 120 \times 120$ instead of the entire volume as a whole. During training we extracted patches with random uniform sampling, while during inference patches are extracted with an overlap of 50% in all the dimensions.

For what concerns the hardware resources, we trained our model in a distributed fashion, exploiting two NVIDIA Quadro RTX 5000 GPUs. The time needed for a complete training is approximately 16h with a batch size of 2.

5.2 Comparison with the State of The Art

In order to compare our proposal with the latest advances in the segmentation of the inferior alveolar nerve [30, 34], Tab. 1 is provided. Both [30] and [34]

Table 1. Comparison of our proposed model with the state of the art on IAC segmentation.

Dataset	Method	IoU	Dice
Maxillo	Usman *et al.* [30]	–	0.770
	Cripriano *et al.* [5]	0.650	0.790
	Zhao *et al.* [34]	–	0.810
	Ours	**0.704**	**0.824**
ToothFairy	Ours	**0.710**	**0.831**

leverage a two-stage approach that aims at filtering out background data before actually performing the canal segmentation. In doing so, [30] makes use of a CNN-based approach that performs worse than both the positional encoding proposed in [5], and the non-deep two-stage approach based on the Frenet frame described in [34]. In Tab. 1 the proposed (complete) model is trained from scratch by means of both 3D "true" label and synthetically generated labels obtained from the deep label expansion phase. For a fair comparison, we performed the training twice, using only the Maxillo dataset (the dataset employed by competitors) and the complete ToothFairy dataset (our reference dataset). The test set of the two datasets matches, being one the extension of the other. The comparative evaluation provided confirms that our proposal outperforms the state-of-the-art competitors on the public dataset, by setting a new upper bound for IAC segmentation.

5.3 On the Effectiveness of the ABS Token and Memory

To showcase the contribution of each model component, we perform our evaluation by progressively including them in Tab. 2. We performed 10 experiments for each setup,[7] but focused only on the deep label expansion phase of the training, thus limiting the number of experiments without losing generality in the conclusion raised (Sec. 4.2). It is worth noticing that any improvement in the deep label expansion step will benefit the whole segmentation pipeline. Moreover, since the model employed in the two phases is the same, the contributions of each proposed component can already be inferred during the generation phase.

At first glance, the comparison between the first two table lines 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.

Introducing the `[ABS]` token to TransPosPadUNet3D (third line of Tab. 2) enhances its performance, already improving with respect to PosPadUNet3D and consistently demonstrating the effect of the proposed ABS token. Furthermore,

[7] Experiments on the same setup differ only in the initialization seed.

Table 2. Contribution of the modules composing our proposal, considering only the generation phase of our training procedure.

Method	Transf.	ABS Token	Memory	Hann Window	Dice
PosPadUNet3D	✗	✗	0	✗	0.797 ± 0.006
TransPosPadUNet3D	✓	✗	0	✗	0.796 ± 0.009
TransPosPadUNet3D	✓	✓	0	✗	0.801 ± 0.005
TransPosPadUNet3D	✓	✗	128	✗	0.800 ± 0.011
TransPosPadUNet3D	✓	✓	128	✗	0.802 ± 0.004
Ours (Complete)	✓	✓	128	✓	$\mathbf{0.809 \pm 0.004}$

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 robustness of the proposed model.

5.4 Qualitative Evaluation

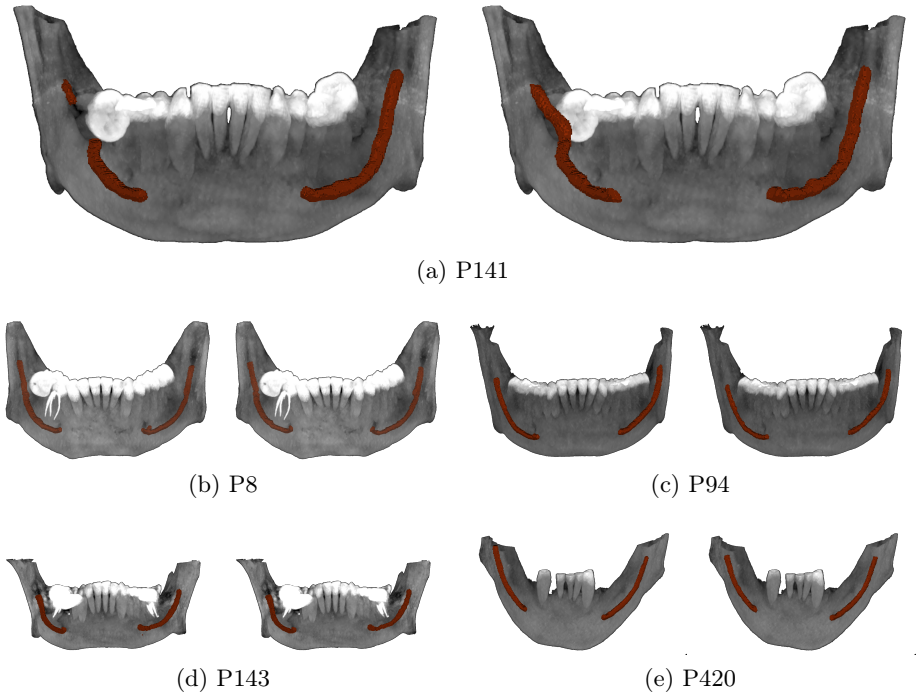
A qualitative evaluation of the predictions obtained using our proposed model is provided in Fig. 6, where five pairs of automatic segmentations are coupled with their corresponding ground-truth annotations. Sample data are taken from the public test case of the ToothFairy dataset.

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 to our model may involve techniques to deal with such a kind of issues.

6 Discussion and Conclusion

One of the primary challenges associated with patch-based learning is the limited context available when modeling patches extracted from the original objects. In order to address this limitation, we propose an innovative approach by incorporating a transformer encoder with memory into the U-Net architecture, along with the introduction of the [ABS] token. Specifically, the [ABS] token is designed to embed the absolute position information of the processed patch within the original volume. By sharing this positional information with other elements within the bottleneck of the U-Net architecture, we are able to enhance the contextual understanding of the patches during the segmentation process and improve overall performance.

Moreover, our transformer encoder is equipped with memory tokens, which serve to store essential and generalized information pertaining to all patches. This



(a) P141

(b) P8 (c) P94

(d) P143 (e) P420

Fig. 6. Segmentation predictions proposed by our model (left) and corresponding ground-truths annotation (right) on examples taken from the ToothFairy public test set. The jaws face the camera view, thus the canal on the left side is the right IAC.

stored information can be particularly valuable for the segmentation task, as it may be difficult to be directly retrieved from each patch singularly. By leveraging the transformer encoder with memory and the [ABS] token, our proposed method seeks to address the contextual information challenge in patch-based learning, improving the segmentation performance within the U-Net architecture.

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 broader 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 critical medical domains to foster progress and innovation.

Future Work. 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 will open doors to a broad range of applications beyond IAC segmentation. This will offer a promising research direction for fur-

ther investigation into its performance across diverse neural networks, datasets, and data modalities.

Acknowledgements. This work was supported by the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2023 and FARD-2024 funds (Fondo di Ateneo per la Ricerca).

References

1. Abdolali, F., Zoroofi, R.A., Abdolali, M., Yokota, F., Otake, Y., Sato, Y.: Automatic segmentation of mandibular canal in cone beam CT images using conditional statistical shape model and fast marching. International Journal of Computer Assisted Radiology and Surgery **12**(4), 581–593 (Apr 2017)
2. Blacher, J., Van DaHuvel, S., Parashar, V., Mitchell, J.C.: Variation in Location of the Mandibular Foramen/Inferior Alveolar Nerve Complex Given Anatomic Landmarks Using Cone-beam Computed Tomographic Scans. Journal of Endodontics **42**(3), 393–396 (2016)
3. Bontempo, G., Porrello, A., Bolelli, F., Calderara, S., Ficarra, E.: DAS-MIL: Distilling Across Scales for MIL classification of histological WSIs. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 248–258. Springer (2023)
4. Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., Minafra, P., Anesi, A., Grana, C.: Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. IEEE Access **10**, 11500–11510 (2022)
5. Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., Grana, C.: Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21137–21146. IEEE (2022)
6. Cornia, M., Baraldi, L., Cucchiara, R.: Explaining transformer-based image captioning models: An empirical analysis. AI Communications **35**(2), 111–129 (2022)
7. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10578–10587 (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D Medical Image Segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 574–584 (2022)
10. Hattab, J., Porrello, A., Romano, A., Rosamilia, A., Ghidini, S., Bernabò, N., Capobianco Dondona, A., Corradi, A., Marruchella, G.: Scoring Enzootic Pneumonia-like Lesions in Slaughtered Pigs: Traditional vs. Artificial-Intelligence-Based Methods. Pathogens **12**(12), 1460 (2023)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

12. Jaskari, J., Sahlsten, J., Järnstedt, J., Mehtonen, H., Karhu, K., Sundqvist, O., Hietanen, A., Varjonen, V., Mattila, V., Kaski, K.: Deep Learning Method for Mandibular Canal Segmentation in Dental Cone Beam Computed Tomography Volumes. Scientific Reports **10**(1), 1–8 (2020)
13. Kainmueller, D., Lamecker, H., Seim, H., Zinser, M., Zachow, S.: Automatic Extraction of Mandibular Nerve and Bone from Cone-Beam CT Data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 76–83. Springer (2009)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems (2012)
15. Landi, F., Baraldi, L., Corsini, M., Cucchiara, R.: Embodied Vision-and-Language Navigation with Dynamic Convolutional Filters. In: Proceedings of the 30th British Machine Vision Conference (2019)
16. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access (2024)
17. Lumetti, L., Pipoli, V., Bolelli, F., Grana, C.: Annotating the Inferior Alveolar Canal: the Ultimate Tool. In: International Conference on Image Analysis and Processing. pp. 525–536. Springer (2023)
18. Mercadante, C., Cipriano, M., Bolelli, F., Pollastri, F., Di Bartolomeo, M., Anesi, A., Grana, C.: A Cone Beam Computed Tomography Annotation Tool for Automatic Detection of the Inferior Alveolar Nerve Canal. In: 16th International Conference on Computer Vision Theory and Applications-VISAPP 2021. vol. 4, pp. 724–731. SciTePress (2021)
19. Moris, B., Claesen, L.J.M., Sun, Y., Politis, C.: Automated tracking of the mandibular canal in CBCT images using matching and multiple hypotheses methods. 2012 Fourth International Conference on Communications and Electronics (ICCE) pp. 327–332 (2012)
20. Pielawski, N., Wählby, C.: Introducing Hann windows for reducing edge-effects in patch-based image segmentation. PloS one **15**(3), e0229839 (2020)
21. Pipoli, V., Cappelli, M., Palladini, A., Peluso, C., Lovino, M., Ficarra, E.: Predicting gene expression levels from dna sequences and post-transcriptional information with transformers. Computer Methods and Programs in Biomedicine **225**, 107035 (2022)
22. Pollastri, F., Cipriano, M., Bolelli, F., Grana, C.: Long-Range 3D Self-Attention for MRI Prostate Segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
23. Pollastri, F., Maroñas, J., Bolelli, F., Ligabue, G., Paredes, R., Magistroni, R., Grana, C.: Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE (2021)
24. Pollastri, F., Parreño, M., Maroñas, J., Bolelli, F., Paredes, R., Ramos, D., Grana, C.: A Deep Analysis on High Resolution Dermoscopic Image Classification. IET Computer Vision **15**(7), 514–526 (October 2021)
25. Porrello, A., Vincenzi, S., Buzzega, P., Calderara, S., Conte, A., Ippoliti, C., Candeloro, L., Di Lorenzo, A., Dondona, A.C.: Spotting Insects from Satellites: Modeling the Presence of Culicoides Imicola Through Deep CNNs. In: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). pp. 159–166. IEEE (2019)

26. Roberti, I., Lovino, M., Di Cataldo, S., Ficarra, E., Urgese, G.: Exploiting Gene Expression Profiles for the Automated Prediction of Connectivity between Brain Regions. International Journal of Molecular Sciences **20**(8), 2035 (2019)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. vol. 9351, pp. 234–241 (2015)
28. Stefanini, M., Lovino, M., Cucchiara, R., Ficarra, E.: Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. Computer Methods and Programs in Biomedicine **234**, 107504 (2023)
29. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
30. Usman, M., Rehman, A., Saleem, A.M., Jawaid, R., Byon, S.S., Kim, S.H., Lee, B.D., Heo, M.S., Shin, Y.G.: Dual-Stage Deeply Supervised Attention-Based Convolutional Neural Networks for Mandibular Canal Segmentation in CBCT Scans. Sensors **22**(24), 9877 (2022)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. Advances in Neural Information Processing Systems (NIPS) **30** (2017)
32. Wei, X., Wang, Y.: Inferior alveolar canal segmentation based on cone-beam computed tomography. Medical Physics (2021)
33. Worthington, P.: Injury of the Inferior Alveolar Nerve during Implant Placement: a Literature Review. International Journal of Oral & Maxillofacial Implants **19**(5) (2004)
34. Zhao, H., Chen, J., Yun, Z., Feng, Q., Zhong, L., Yang, W.: Whole mandibular canal segmentation using transformed dental CBCT volume in Frenet frame. Heliyon **9**(7) (2023)