Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation

Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana
Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it

Abstract

Many recent works in dentistry and maxillofacial imagery focused on the Inferior Alveolar Nerve (IAN) canal detection. Unfortunately, the small extent of available 3D maxillofacial datasets has strongly limited the performance of deep learning-based techniques. On the other hand, a huge amount of sparsely annotated data is produced every day from the regular procedures in the maxillofacial practice. Despite the amount of sparsely labeled images being significant, the adoption of those data still raises an open problem. Indeed, the deep learning approach frames the presence of dense annotations as a crucial factor. Recent efforts in literature have hence focused on developing label propagation techniques to expand sparse annotations into dense labels. However, the proposed methods proved only marginally effective for the purpose of segmenting the alveolar nerve in CBCT scans. This paper exploits and publicly releases a new 3D densely annotated dataset, through which we are able to train a deep label propagation model which obtains better results than those available in literature. By combining a segmentation model trained on the 3D annotated data and label propagation, we significantly improve the state of the art in the Inferior Alveolar Nerve segmentation.

1. Introduction

Dental implant placement within the jawbone is a routinely executed surgical procedure, which can become complex due to the local presence of the Inferior Alveolar Nerve (IAN). In particular, the nerve is oftentimes in close relation to the roots of molars, and its position must thus be carefully detailed before the surgical removal. As avoiding contact with the IAN is a primary concern during these operations, segmentation plays a key role in surgical preparations.

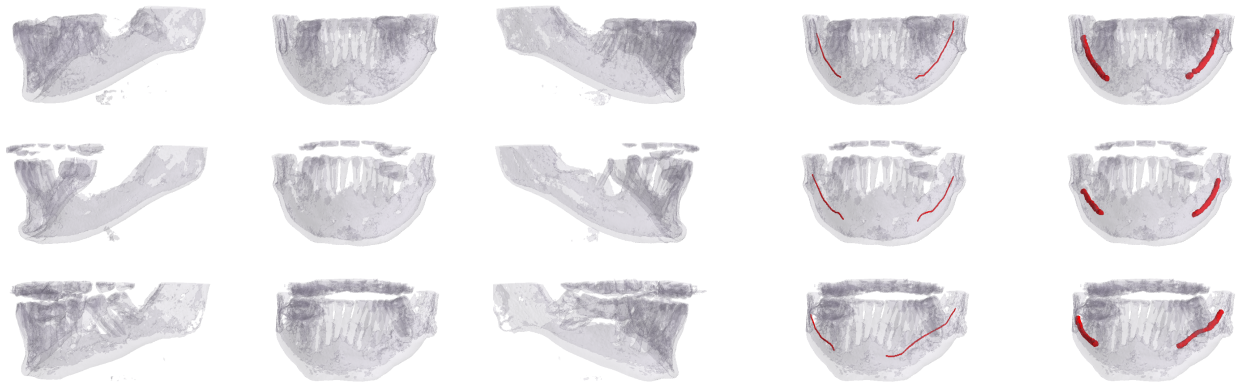
Given the exceptionally large amount of time required for 3D manual segmentation, perfect anatomical annotation accuracy is usually overlooked in favour of a fast execution time. Therefore, the *de facto* standard in radiology medical centers for dentistry and maxillofacial purposes is based on sparse annotations, which can be obtained from 2D images in a relatively small amount of time. Nevertheless, 2D annotations fail to identify a considerable amount of inner information about the IAN position and the bone structure. The incomplete detection of the nerve positioning is often sufficient to facilitate a positive outcome of surgical intervention, but it is not an accurate anatomical representation.

Convolutional Neural Networks (CNNs) have provided amazing results for both 2D and 3D segmentation, alongside several more computer vision tasks [2, 5, 9, 11, 13, 26]. As a matter of fact, a segmentation CNN would be able to correctly portray the 3D structures of an IAN without any need for manual adjustment. Unfortunately, the great capabilities of CNNs in this field are limited by the lack of carefully annotated data, which is indispensable for training deep learning models. Indeed, despite the significant amount of raw data available, the supervised learning paradigm requires dense 3D annotations to reach its full potential, which are extremely expensive to acquire.

In this work, we propose a novel label propagation method, based on deep learning, that can translate sparse 2D labels into 3D voxel-level annotations. This method can fill the gap between the most modern and sophisticated methods for 3D segmentation and the lack of viable annotated data in the maxillofacial field. Moreover, with the goal of pushing the state of the art in 3D IAN segmentation, we design a novel 3D segmentation CNN that exploits positional information to generate the final 3D prediction.

In order to correctly evaluate both of the proposed methods and any future competitor, a new dataset has been collected, annotated by medical experts at voxel-level (Fig. 1e), and publicly released along with this paper[1]. We address this as *3D annotation* in contrast to the traditional one, which is performed on a 2D "panoramic view" obtained from the volume (Fig. 1d).
The main contributions introduced by this paper can be

[1] https://ditto.ing.unimore.it/maxillo/



(a) (b) (c) (d) (e)

Figure 1. Samples from the proposed dataset. Each line of the image contain a different patient, from left to right you can see (a) left-side, (b) frontal and (c) right-side views of the CBCT volume. (d) and (e) depicts sparse and dense annotations of the inferior alveolar nerve respectively. Best viewed in color.

summarized as follows:

- We design a novel deep label propagation technique to enhance sparse 2D annotations and yield dense voxel-level annotations;
- A novel deep learning architecture for 3D segmentation is proposed, which improves the state-of-the-art segmentation accuracy for the mandibular canal;
- We collect and publicly release the first CBCT (Cone Beam Computed Tomography) 3D dataset with professionally produced 3D annotations;
- The source code which allows to exactly reproduce all the reported experiments is publicly released.

The rest of the paper is organized as follows. Sec. 2 presents state-of-the-art approaches for both the automatic detection of the inferior alveolar nerve and label propagation. Sec. 3 and Sec. 4 describe respectively the collected dataset and the proposed methods, both for label propagation and 3D IAN segmentation. Finally, experiments are detailed in Sec. 5, and conclusions are drawn in Sec. 6.

2. Related Works

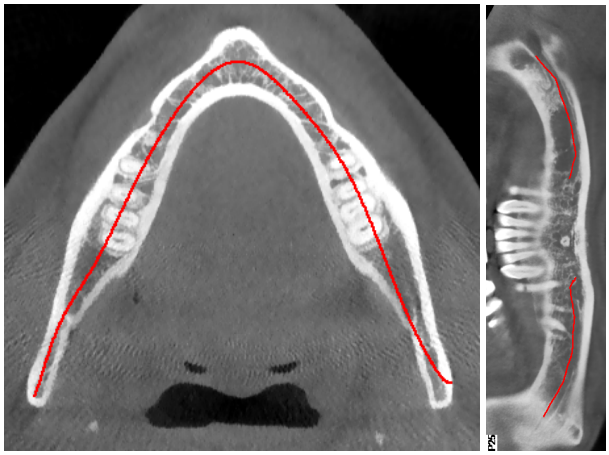
2.1. IAN Segmentation

Since the early 2000s, the worldwide spread of Cone Beam Computed Tomography (CBCT) [28] has brought the scientific community to devote many efforts to the development of automatic systems for the segmentation of the IAN in CBCT scans, using classical computer vision methods at first [1, 16, 17, 23, 32], and machine learning and deep learning more recently [6, 14, 15, 18]. Classical computer vision methods are mostly based on Statistical Shape Model (SSM), and enhanced by means of either tracing or fast marching algorithms [1, 16]. These methods, however, are limited by the need for segmented mandible bone in the training annotation, which requires additional manual work. As an alternative, predefined thresholds can be used to separate tissue from the canal [23], but this often results in an incomplete depiction because of the low contrast characterizing CBCT scans.

In 2020, Jaskari *et al*. [15] designed one of the first deep learning applications to the segmentation of the mandibular canal, by training a fully convolutional network on a dataset of coarsely annotated 3D scans. Their annotations are obtained by manually selecting an average of 10 control points for each canal, interpolating them into a spline, and applying a static 3 mm diameter to the outline. Their approach achieves better results than previous attempts based on SSM, but the ground truth masks automatically generated from coarse annotations hamper the performance of the method. In the same year, Kwak *et al*. [18] trained 2D and 3D models based on SegNet [2] and U-Net [5, 27] using a private annotated dataset. However, neither the dataset nor the code is publicly available, making it impossible to compare their approach to ours. Our work focuses on the 3D inferior alveolar canal segmentation, proposing a new architecture, a public dataset, and a novel deep label propagation technique. The full dataset with annotations and subdivision in training and testing splits are available in [7], together with the code to reproduce the experiments proposed in this paper.

2.2. Label Propagation

Because of the heavy burden represented by manual annotation of segmentation ground truth, researchers have recently focused on the development of models that exploit sparse annotations instead, in different styles, such as scribbles, bounding boxes, or points [3, 8, 12, 20, 21]. In particular, the use of scribble annotations in the field of image segmentation has been recently investigated in [19, 30, 31, 33],



(a) Axial Slice (b) Panoramic

Figure 2. (a) axial slice. The panoramic base curve (red line) identifies the mandible and is used to generate the panoramic view (b) to produce the final 2D annotation. Best viewed in color.

where different methods are proposed to propagate sparse labels and thus produce dense ground truths.

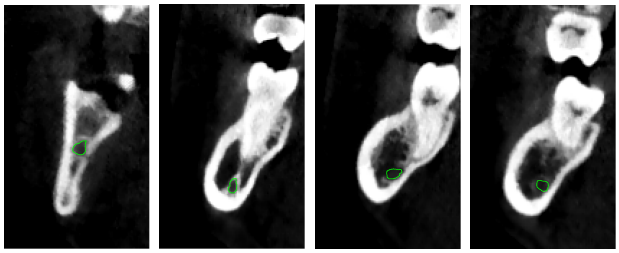
Fewer proposals are related to the medical imaging field. In [25], a method to recover semantic segmentations given a database of brain and lung scans with corresponding bounding boxes is proposed, which makes use of an iterative energy minimization problem defined over a densely connected Conditional Random Field (CRF) to update the parameters of a CNN model. Unfortunately, the approach is limited to 2D, and bounding boxes do not suit well the concave and diagonal extension of the IAN.

The 3D version of U-Net [5] was originally introduced with the aim of learning to perform 3D dense segmentation from sparse annotations. The network is trained using a few annotated slices as ground truth, and leaving most voxels unlabeled. The method is still constrained by the need for thoroughly annotated slices, which require considerably more manual work than simply drawing a pair of lines.

Regarding the specific topic of this paper, the segmentation of the IAN, the only label propagation proposal up to now is represented by the already mentioned work by Jaskari *et al*. [15], where sparse annotations in the form of lines are propagated to produce a tubular shape that mimics the body of the mandibular canal which contains the nerve.

3. Dataset

State-of-the-art algorithms for inferior alveolar nerve canal segmentation are based on annotations obtained from 2D views. In fact, the huge amount of time required to realize 3D annotations makes their cost prohibitive for most medical centers. Furthermore, the datasets used to produce the current state-of-the-art results are private and not accessible to the scientific community. This prevents researchers



(a) (b) (c) (d)

Figure 3. Example of Cross-Sectional Views (CSVs) with annotations of the mandibular canal. Best viewed in color.

from replicating experiments and validating their novelties.

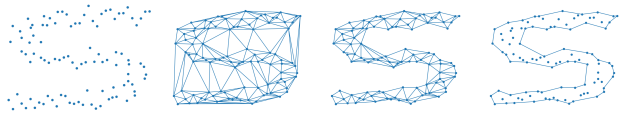
Such kind of 2D annotations identify the upper part of the mandibular canal along the entire dental arch. The annotation process, usually performed by radiology technicians, is divided into three different steps. An axial plane of the original CBCT scan (Fig. 2a) is selected and a spline —the panoramic base curve— is manually drawn to fit the central part of the mandible. This spline identifies the panoramic view, *i.e*., a 2D image constituted by the voxels of the curved plane orthogonal to the axial slice and crossing the panoramic base curve. This view highlights the inferior alveolar nerve and is employed by radiologists to produce the sparse annotation of the IAN (Fig. 2b). The resulting annotation can be mapped back to the original 3D volume; examples are reported in Fig. 1d. In the rest of the discussion we will refer to 2D annotations as *sparse*, in contrast with 3D *dense* annotation described below. Sparse 2D annotations are employed in everyday surgical practice to measure the height and depth of artificial implants and avoid inferior alveolar nerve injuries.

To cope with the lack in literature reported at the beginning of this Section, this paper gathers and publicly releases a new dataset of 3D CBCT scans with *dense* 3D annotations of the inferior alveolar nerve. The dataset counts 347 dental scans obtained by means of CBCT (*NewTom/NTVGiMK4*, 3 mA, 110 kV, 0.3 mm cubic voxels). Volumes have been acquired with the 0.3 mm intra-slice distance with a shape in the range from $(148, 265, 312)$ to $(178, 423, 463)$ for the Z, Y and X axes. Voxel values, expressed in Hounsfield unit (HU), are in the interval $[-1\,000, 5\,264]$. Sparse annotations are available for all of the 347 volumes composing the proposed dataset, and 91 volumes have been elaborated by a team of doctors with years of experience in maxillofacial surgery to produce *dense* voxel-level annotations of the canal. The rest of the paper will refer to the set 91 volumes with both dense e sparse annotation as *primary* dataset, whereas we will identify the 256 sparsely annotated volumes as *secondary* dataset.

The annotation procedure employed by expert doctors to produce the final labels is summarized below and has been carried out by means of the tool described in [22]. First of

(a) Dense point set (b) α-shape mesh (c) Voxelized volume

Figure 4. Dense annotations pre-processing. The annotations produced as described in Sec. 3 are dense and jagged (a). For this reason we compute a concave α-shape (b). (c) is the resulting binary raster volume after voxelization.



(a) Point Set (b) Triangulation (c) α-complex (d) α-shape

Figure 5. α-shape construction process for the point set of (a). In the Delaunay triangulation (b), triangles with circumradius $\leq -\frac{1}{\alpha}$ form a simplicial subcomplex known as α-complex (c), whose border is the α-shape (d).

all, the arch approximation that better describes the canal course is automatically identified through morphological operations and possibly manually adjusted (Fig. 2a). The one-voxel thick curve produced as output is then approximated with a polynomial and converted into a Catmull-Rom spline. Each point of the spline generates a *Cross-Sectional Line* (CSL) lying on the axial plane and perpendicular to the spline itself. Starting from the CSL, a *Multi Planar Reformation* (MPR) is performed in order to generate a *Cross-Sectional View* (CSVs). The process consists in interpolating the value of the base lines (CSL) across the entire volume height. The CSVs can be optionally rotated to be orthogonal to the canal slope, thus ensuring the canal to be circular in each view and simplifying the following annotation process which is preformed drawing closed Catmull-Rom splines on these views (Fig. 3). The ground truth volumes constituting the dataset are generated refining the set of points filling the closed splines by means of the α-shape algorithm detailed in the following.

3.1. Alpha-Shape Annotation Refinement

The annotation tool used for producing dense segmentation ground truths usually outputs a jagged point set, similar to the example of Fig. 4a. In order to obtain a smooth polygonal mesh out of it, we compute a concave α-*shape*.

The α-shape [10] is a generalization of the convex hull, aimed at representing the intuitive concept of shape of a point set. The only parameter of the algorithms is $\alpha \in \mathbb{R}$, that regulates the "crudeness" of the result. Eq. (1) defines a *generalized disk of radius* $\frac{1}{\alpha}$, D_α:

$$D_\alpha \begin{cases} \text{The complement of a disc of radius } -\frac{1}{\alpha}, & \text{if } \alpha < 0 \\ \text{A halfplane,} & \text{if } \alpha = 0 \\ \text{A disc of radius } \frac{1}{\alpha}, & \text{if } \alpha > 0 \end{cases} \quad (1)$$

Given a point set S and a value for α, the α-shape is constructed in this way: an edge is put between two points p_i and p_j whenever there exists a D_α with p_i and p_j lying on its boundary, and which contains the entire S. When $\alpha = 0$, this procedure constructs the convex hull; instead, cruder or finer shapes can be respectively obtained using positive or negative α values. Because of the geometrical nature of the alveolar nerve, we are specifically interested in concave α-shapes, achievable when $\alpha < 0$.

The most common method for computing a concave α-shape consists of taking the border of a simplicial subcomplex extracted from the Delaunay triangulation, containing only triangles with circumradius $\leq -\frac{1}{\alpha}$. An example of the process is depicted in Fig. 5.

The above concepts can be extended to the three-dimensional case by substituting disks and triangles with spheres and tetrahedra. Fig. 4b illustrates an example of α-shape polygonal mesh built starting from dense annotations of the alveolar nerve. Finally, the mesh is converted into a binary raster volume by means of voxelization: the final result is given in Fig. 4c.

4. Method

This Section describes the label propagation approaches for the secondary sparsely annotated dataset and the models for the main task, namely the deep segmentation of the mandibular canal from CBCT images. We firstly report our competitor methodologies and then detail our proposals.

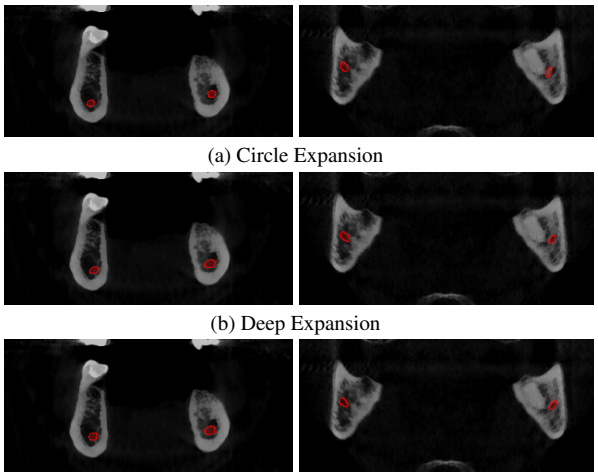
4.1. Reference Approach

The work by Jaskari *et al*. [15] introduced the following fully automatic method to obtain synthetic dense labels from sparse annotations. For each annotation point, the direction of the canal is determined using the coordinates of the next point. Then, a circle with a diameter of 3 millimeters is drawn on the plane orthogonal to the canal direction. Finally, all the circles are connected in a hollow pipe-shaped 3D structure, that is voxelized and filled with traditional computer vision algorithms.

This method is employed by the authors to construct their training set. In the following, this kind of synthetic labels will be referred to as *Circle Expansion*, in opposition to the new densely annotated dataset and our novel Deep Expansion technique, both introduced with the present work. The training set is prepared offline. Input values, stored in Hounsfield unit (HU), are cropped to avoid peak values caused by metal artifacts in the patient mouth or acquisition noise. A grid sampling [24] is run to extract 32^3-sized patches, and patches without any mandibular canal voxel are discarded to reduce class imbalance. The segmentation method proposed by Jaskari *et al*. makes use of a custom version of 3D U-Net, with short residual connections

(a) (b) (c) (d) (e)

Figure 6. (a) is the ground truth. Competitor model prediction before (b) and after (c) noise filtering. Finally, the last two figures depict the prediction after training our architecture on the primary dataset (d) and when the pre-training on the dataset generate through the label-propagation network is included (e).



(a) Circle Expansion

(b) Deep Expansion

(c) Manual 3D Annotation

Figure 7. Cross-sectional views at different z-depth. Circle Expansion (a), our novel Deep Expansion network (b), and hand-drawn dense annotations (c) are compared. While the limits of the circle expansion approach are blatant, the prediction of our deep learning model (b) is very close to the 3D manual annotation (c).

between blocks with the same feature channels, a different depth —maximum number of channels is 256 instead of 512— and convolutions with a stride of two in order to avoid max pooling layers. The 3D network outputs are refined by a post-processing which only keeps the two largest connected components [4], with the aim of removing false positive voxels Fig. 6c.

Since no public source code of the method is available, our implementation is provided, in order to allow the reader to completely reproduce the results reported in Sec. 5.

4.2. Proposed Method

We employ a modified version of 3D U-Net as a backbone for both the label propagation and the IAN segmentation CNNs. All of the volumes employed in this work have 0.3 voxel spacing, *i.e*. the same resolution despite the different overall dimensions. Therefore, we extract the center of every volume without ever discarding parts of the annotated canal and obtaining sub-volumes with a size of $168 \times 280 \times 360$ voxels. During training, sub-volumes are further divided in a grid of $80 \times 80 \times 80$ blocks.

Differently from the original 3D U-Net architecture, every three-dimensional convolution in our CNN applies a 2 pixels padding along each dimension. Although this alteration does not cause any variation in terms of performance, it ensures that the output of each convolution has the same size of its input along axis x, y, and z.

Resolution changes are therefore due uniquely to the three max pooling layers, each halving the size of the volumes. When using the aforementioned input size, the encoder output is composed of 512 feature maps of size $10 \times 10 \times 10$. Inside the decoder, on the other hand, resolution changes are caused by transposed convolutions, with dimensions $2 \times 2 \times 2$ for both the kernel and stride to double the size of the feature maps. This adjustment ensures resolution symmetry between features in the decoder and corresponding maps in the encoder, thus allowing them to be simply concatenated with skip connections. Moreover, the output of the model naturally has the same dimensions of the input. The final output is a single channel volume, to which we apply a Sigmoid activation function and a threshold at 0.5 to obtain the final binary prediction mask.

The segmentation architecture is further enriched with a *positional embedding*. Since our sub-volumes are extracted from the original scan following a fixed grid, we exploit positional information derived from the location of the top-left and bottom-right corners of the sub-volume. Specifically, these global coordinates are fed to a linear layer which yields a single feature map of dimensions $10 \times 10 \times 10$. This positional embedding gets concatenated to the output of the encoder, and then fed to the decoder. Exploiting positional information ensures two extremely important benefits:

- During training, the CNN is fed with implicit information about areas close to the edges of the scan, where the IAN is very unlikely to be present. This piece of knowledge greatly reduces the number of false positives during inference. As a matter of fact, no post-processing method is required to refine the output of the proposed CNN, contrarily to [15].
- Information about cut positions helps the network to better shape the output: sub-volumes located close to the mental foramen generally present a much thinner canal than those located in the mandibular foramen.

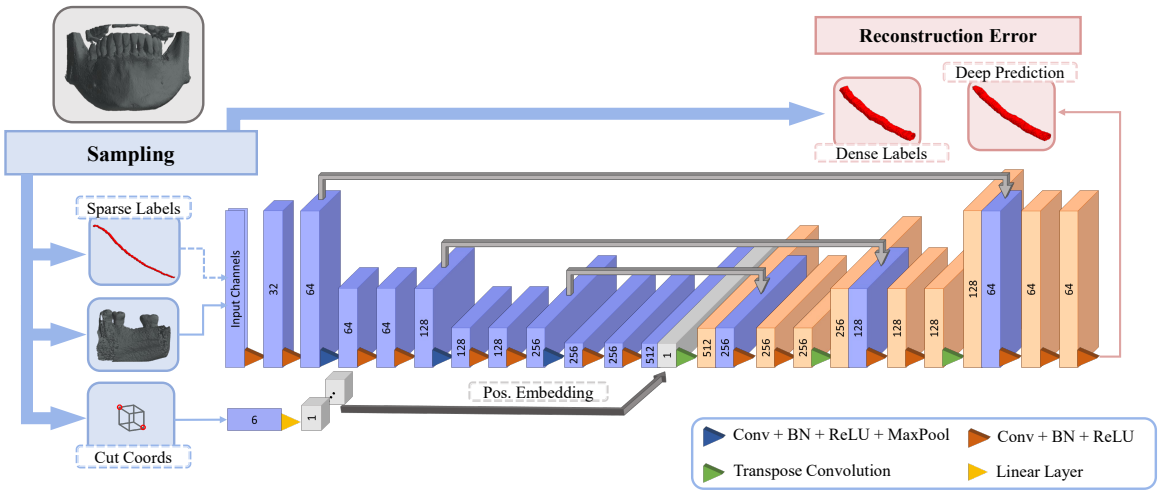


Figure 8. Architecture of the proposed Positional PadUNet. The 3D CNN has a classic encoder-decoder structure, and the output is a single-channel prediction mask. The input is a 3D sub-volume obtained from a CBCT scan, enriched with information about its global location. As an approach for the label propagation task, sparse annotations are concatenated with the input volume. Best viewed in color.

This technique could indeed be employed for several classes of medical data, as anatomical structures can substantially vary according to their location.

We name the proposed architecture *Positional PadUNet*. A detailed description is provided in Tab. 1.

4.2.1 Deep Expansion

The presence of a new 3D voxel-level annotated dataset paves the way to a new frontier for label propagation: we can indeed employ 3D annotations to supervise a deep label propagation neural network, trained to expand sparse labels into dense ones, and produce high quality synthetic ground truths for the segmentation task. We christen this new label propagation approach *Deep Label Expansion*, or *Deep Expansion* in short. The deep expansion model is based on the proposed segmentation network, Positional PadUNet. The main difference regards the input layer, that is changed in order to accept a concatenation of both the raw volume data and the sparse annotations, rendered as a binary channel. Thus, the network input has a shape of $2 \times 80 \times 80 \times 80$. Same as Positional PadUNet, a positional embedding is concatenated to the encoder output. By means of this model, we generate a synthetic dataset which will be employed to pre-train our final segmentation CNN. Once again, the positional embedding supplies important information about the location of the cut, which is closely related to the diameter of the expanded labeled canal.

Layer	Input Channels	Output Channels	Skip Connections
3D Conv Block 0	1 or 2	32	✗
3D Conv Block 1	32	64	✓
3D Conv Block 2	64	64	✗
3D Conv Block 3	64	128	✓
3D Conv Block 4	128	128	✗
3D Conv Block 5	128	256	✓
3D Conv Block 6	256	256	✗
3D Conv Block 7	256	512	✗
Transpose Conv	513	512	✗
3D Conv Block 8	512 + 256	256	✓
3D Conv Block 9	256	256	✗
Transpose Conv	256	256	✗
3D Conv Block 10	256 + 128	128	✓
3D Conv Block 11	128	128	✗
Transpose Conv	128	128	✗
3D Conv Block 12	128 + 64	64	✓
3D Conv Block 13	64	64	✗
3D Conv Block 14	64	1	✗

Table 1. Overview of our backbone for both the proposed deep label expansion network (two input channels) and segmentation network (one input channel). Every layer from the encoder that produces skip connections is followed by a 3D max pooling layer with kernel size 2. All the *3D Conv Blocks* are made out of a 3D convolutional layer with kernel size 3, stride 1 and padding 1, followed by a 3D Batch Normalization layer and a ReLu activation function. Every Transpose Convolution has kernel size and padding of two.

5. Experimental Results

This Section firstly describe the improvements we introduced in the competitor pipeline (Sec. 5.1), we will then evaluate the benefit provided by the proposed densely annotated 3D dataset, observing a ground-breaking improvement in the network accuracy (Sec. 5.2). The benefit of the proposed deep expansion model are finally highlighted in

# Test	Model	Pre-Training Set	Training Set	Sampling	Batch Size	Vol Shape	IoU	Dice
1	Jaskari *et al*.	-	Cir.Exp.	Grid	24	32^3	0.39	0.56
2	Jaskari *et al*.	-	Cir.Exp.	Weighted Grid	6	80^3	0.44	0.61
3	PadUNet*	-	Cir.Exp.	Weighted Grid	6	80^3	0.48	0.64
4	Pos.PadUNet*	-	Cir.Exp.	Weighted Grid	6	80^3	0.48	0.65
5	PadUNet*	-	3D Ann.	Weighted Grid	6	80^3	0.58	0.73
6	Pos.PadUNet*	-	3D Ann.	Weighted Grid	6	80^3	0.61	0.75
7	Pos.PadUNet*	Cir.Exp.	3D Ann.	Weighted Grid	6	80^3	0.63	0.77
8	Pos.PadUNet*	Deep Expansion*	3D Ann.	Weighted Grid	6	80^3	**0.65**	**0.79**

Table 2. Experimental results using different datasets and methods. Novel models and techniques are marked with a star.

Dataset Name	Split	Sparse	Dense	n. Volumes
Primary	Training	✓	✓	68
	Validation	✓	✓	8
	Testing	✓	✓	15
Secondary	Pre-Training	✓	✗	256

Table 3. Summary of the datasets employed for the experiments reported in Sec. 5.

Sec. 5.3, clearly demonstrating the effectiveness of our positional embedding.

Before getting to the heart of the discussion, it is important to underline that during inference —contrary to the training phase— volumes are not center cropped, *i.e*. each part of the volume is fed to the model. Moreover, both the deep expansion and the segmentation network are tested on the same data split, thus ensuring a proper and fair comparison. Indeed, employing a single split for both tasks ensures that the synthetic training set does not implicitly contain important features that characterize samples from the test set.

During training, the following 3D augmentations are employed: (i) random affine transformations with scaling between 0.5 and 1.5, and rotation between -10 and 10 degrees with probability 0.5; (ii) random flip on the x axis with probability 0.7 (canal branches tend to be symmetric).

Although both Intersection over Union and Sørensen–Dice similarity [29] (Dice score) are presented, their values are very strictly correlated, and thus only the latter is discussed in the following.

5.1. Improving Existing Method

In order to fairly compare with [15], the first experiment trains our implementation of their pipeline with our primary dataset (Tab. 3). Although the number of volumes is slightly lower, the final results (line #1 of Tab. 2) is practically the same (0.57 vs 0.56 of Dice score) after applying their post-processing technique thus ensuring the correctness of our understanding and implementation.

To reproduce their results it is mandatory to apply the post-processing method, to cope with the large number of false-positive voxels predicted by the network (Fig. 6b). As the authors suggest in the paper, this output noise is likely related to the sampling approach: discarding sub-volumes with no ground truth voxels causes the network to always expect input patches to contain part of the foreground.

By introducing a *weighted grid sampling* process we are able to eliminate any artifact produced by the network, removing the need of any post-processing. In weighted grid sampling, 3D patches are generated online. Each patch inside the batch is weighted as the number of foreground voxels it contains over the maximum number of foreground voxels in any patch of the current batch. Weight values are then provided to the loss function to balance the relevance of each patch volume according to its number of ground truth voxels.

Differently from the original proposal, only batches entirely composed of background sub-volumes are discarded, meaning we allow the network to see "empty" patches. This sampling method ensures the network is fed with almost any part of the original input data, while the weighting system mitigates the risk of predicting all-background volumes. Additionally, we improve predictions of the model by enlarging patch volumes at the cost of reducing the batch size from 24 to 6. These changes give the network a more global view of the spatial features and significantly increase the quality of the final results reported in line #2 of Tab. 2.

Thanks to these changes, the model achieves an improvement of five Dice points without requiring any post-processing to clean up the raw network output, preserving the end-to-end training.

5.2. Improving the Model and the Data

Introducing the model described in Sec. 4.2 in addition to the adjustments described above, it is possible to further achieve some improvements. PosPadUNet provides more channels, uses max pooling instead of strided convolutions, removes the element-wise summations, and includes a positional encoding of the 3D patch, for a 0.04 Dice score increment (line #4 of Tab. 2).

By training PosPadUNet on our dense voxel-level annotations dataset, we achieve an impressive 0.75 Dice score,

19 points higher than the current SOTA. The same model trained on densely annotated data gives 10 points of improvement in the Dice metric with respect to using a synthetically expanded dataset (line #6 vs #4 of Tab. 2). This result proves how the new dataset is indeed suitable and effective for deep learning purposes and highlights the strong limitation of a naive label propagation method.

The impact of the proposed positional encoding is evaluated by looking at lines #3 and #5. Here, we remove any kind of location information from the input of the network, thus slightly simplifying the architecture depicted in Fig. 8. The performance of this downgraded version of the proposed segmentation network, titled PadUNet in contrast to the whole proposed architecture, shows how enriching the network input with knowledge about the original position of the cut is more effective on a dense annotation setup *w.r.t.* the network trained using circle expansions as ground truth. To our understanding, this is related to the ability of voxel-level labels to correctly represent the correlation between the inferior alveolar canal diameter and its location within the mandible, whereas circle expansions ignore the tendency of the canal to thicken when getting closer to the mental foramen.

5.3. Deep Label Expansion

The deep label expansion model described in Sec. 4 is employed to generate 256 3D synthetic voxel-level annotations for the volumes with sparse annotations only, *i.e.* the secondary dataset. The quality of these synthetic labels composing the so-called deep expansion dataset is assessed in Tab. 4, by means of the same two metrics adopted in the rest of this Section. The first two rows compare our label propagation method with the circle expansion algorithm proposed by Jaskari *et al.* (Sec. 4.1), making use of the proposed test set annotated at a voxel level. Experimental results clearly demonstrate an accuracy improvement.

The last row of Tab. 4, on the other hand, depicts the performance of Positional PadUNet when trained for the IAN segmentation task, without making use of the 2D sparse manual annotations. This last row has the purpose of proving how making use of coarse annotations actually improves the quality of the final output. It is very important to highlight that the first row of Tab. 4 evaluates the accuracy of labels obtained through circular expansion, whereas the first four rows of Tab. 2 assess the performance of CNNs trained with such automatically generated annotations.

The deep expansion dataset can be used for pre-training our segmentation network, which is subsequently fine-tuned on the 3D manually annotated dataset (line #8 of Tab. 2). As already reported, without any fine-tuning set-up, the model achieves a 0.75 Dice score on the segmentation of the IAN. The deep expansion dataset allows to obtain a 0.04 improvement in terms of Dice score. As a comparison, performing a pre-training on the circle expansion dataset only increases the Dice score by 0.02.

Model	Data	Use Sparse	IoU	Dice
Circle Expansion	-	✓	0.38	0.55
Deep Expansion	3D	✓	**0.64**	**0.78**
Deep Expansion	3D	✗	0.61	0.75

Table 4. Experimental results when training our deep label propagation network compared to the baseline results on our densely annotated testset.

# Test	IoU ($\pm$ std)	Dice ($\pm$ std)
6	0.60 ($\pm$0.014)	0.75 ($\pm$0.013)
7	0.62 ($\pm$0.021)	0.76 ($\pm$0.017)
8	0.64 ($\pm$0.017)	0.78 ($\pm$0.023)

Table 5. Experimental results employing different dataset splits.

Hence, combining all of the novelties, both the new 3D training dataset and the novel deep expansion network, we are able to reach a final Dice score of 0.79; a 41% improvement *w.r.t.* our competitor. This is by far the highest score ever obtained in the task and the new state of the art for the inferior alveolar canal segmentation.

5.4. Improving the Evaluation Protocol

In order to establish a solid benchmark for the segmentation of the IAN, we replicate experiments #6, #7, and #8 of Tab. 2 by using six ($91//15 = 6$) different random dataset splits and ensuring that test sets are disjoint. This allows us to calculate the mean and the standard deviation of both IoU and Dice metrics (Tab. 5), confirming that the obtained results are independent from the split.

6. Conclusions

With this paper we tackled the 3D segmentation of the Inferior Alveolar Nerve. We focused on the importance of voxel-level annotations for the task, gathering a new dataset with finely-grained labels and proposing a novel label propagation method that allows 3D dense annotations to be generated from 2D outlines. Experimental results confirmed both the great relevance of dense high quality 3D labels, and the utility of enhancing 2D annotations with the proposed approach. Moreover, the novel segmentation CNN presented in this work pushes the state of the art in IAN segmentation to a 0.79 Dice score. To the best of our knowledge, this is the first public maxillofacial dataset with voxel-level annotations of the Inferior Alveolar Nerve. Our work aims to encourage the scientific community to further improve results for the IAN canal segmentation tasks, by distributing a public dataset feasible for deep learning training. The new collected dataset and the methods described in this paper are publicly available in [7].

References

[1] Fatemeh Abdolali, Reza Aghaeizadeh Zoroofi, Maryam Abdolali, Futoshi Yokota, Yoshito Otake, and Yoshinobu Sato. 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

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2017. 1, 2

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the Point: Semantic Segmentation with Point Supervision. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 549–565, Cham, 2016. Springer International Publishing. 2

[4] Federico Bolelli, Stefano Allegretti, and Costantino Grana. One DAG to Rule Them All. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2021. 5

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016. 1, 2, 3

[6] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Mattia Di Bartolomeo, Federico Pollastri, Arrigo Pellacani, Paolo Minafra, Alexandre Anesi, and Costantino Grana. Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. *IEEE Access*, 10:11500–11510, 2022. 2

[7] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana. Paper Source Code & Dataset. https://ditto.ing.unimore.it/maxillo/. Accessed: 2022-03-28. 2, 8

[8] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 1

[10] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the Shape of a Set of Points in the Plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. 4

[11] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. 1

[12] Weifeng Ge, Sheng Guo, Weilin Huang, and Matthew R Scott. Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3345–3354, 2019. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] Jae-Joon Hwang, Yun-Hoa Jung, Bong-Hae Cho, and Min-Suk Heo. An overview of deep learning in the field of dentistry. *Imaging science in dentistry*, 49(1):1–7, 2019. 2

[15] Joel Jaskari, Jaakko Sahlsten, Jorma Järnstedt, Helena Mehtonen, Kalle Karhu, Osku Sundqvist, Ari Hietanen, Vesa Varjonen, Vesa Mattila, and Kimmo Kaski. Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes. *Scientific reports*, 10(1):1–8, 2020. 2, 3, 4, 5, 7

[16] Dagmar Kainmueller, Hans Lamecker, Heiko Seim, Max Zinser, and Stefan Zachow. Automatic extraction of mandibular nerve and bone from cone-beam ct data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 76–83. Springer, 2009. 2

[17] Dirk-Jan Kroon. *Segmentation of the Mandibular Canal in Cone-Beam CT Data*. PhD thesis, University of Twente, Netherlands, Dec. 2011. 10.3990/1.9789036532808. 2

[18] Hyunjung Kwak, Eun-Jung Kwak, Jae-Min Song, Hae Park, Yun-Hoa Jung, Bong-Hae Cho, Pan Hui, and Jae Hwang. Automatic mandibular canal detection using a deep convolutional neural network. *Scientific Reports*, 10, 12 2020. 2

[19] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016. 2

[20] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One Thing One Click: A Self-Training Approach for Weakly Supervised 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1726–1736, June 2021. 2

[21] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep Extreme Cut: From Extreme Points to Object Segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 616–625, 2018. 2

[22] Cristian Mercadante, Marco Cipriano, Federico Bolelli, Federico Pollastri, Alexandre Anesi, Costantino Grana, et al. 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*, 2020. 3

[23] Behnam Moris, Luc J. M. Claesen, Yi Sun, and Constantinus Politis. Automated tracking of the mandibular canal in CBCT images using matching and multiple hypotheses methods. *2012 Fourth International Conference on Communications and Electronics (ICCE)*, pages 327–332, 2012. 2

[24] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, page 106236, 2021. 4

[25] Martin Rajchl, M. J. Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Bernhard Kainz, and Daniel Rueckert. DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 36:674–683, 2017. 3

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Nov 2015. 2

[28] William C. Scarfe, Allan G. Farman, and P. Sukovic. Clinical Applications of Cone-Beam Computed Tomography in Dental Practice. *Journal of the Canadian Dental Association*, 72 1:75–80, 2006. 2

[29] Th A Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34, 1948. 7

[30] Paul Vernaza and Manmohan Chandraker. Learning Random-Walk Label Propagation for Weakly-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7158–7166, 2017. 2

[31] Chun Wang, Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, and Chengqi Zhang. Attributed Graph Clustering: A Deep Attentional Embedding Approach. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3670–3676. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 2

[32] Xueqiong Wei and Yuanjun Wang. Inferior alveolar canal segmentation based on cone-beam computed tomography. *Medical Physics*, 2021. 2

[33] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-Supervised Salient Object Detection via Scribble Annotations. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12543–12552, 2020. 2