Automatic teeth segmentation on panoramic X-rays using deep neural networks

Rafic Nader, Andrey Smorodin, Natalia de La Fourniere, Yves Amouriq, Florent Autrusseau

Automatic teeth segmentation on panoramic X-rays using deep neural networks

Rafic Nader
LTeN, Polytech'Nantes (U6607), and
University of Nantes, France
Rafic.Nader@univ-nantes.fr

Andrey Smorodin
Odessa Polytechnic National University
Ukraine, Odessa, Shevchenko ave., 1.
andrey.v.smorodin@op.edu.ua

Natalia De La Fournière
Artefakt-AI, 53 rue Felix Thomas,
44000 Nantes France
ndelafourniere@artefakt-ai.com

Yves Amouriq
UFR Odontologie,
RMeS lab. Inserm UMR-1229
University of Nantes, France
Yves.Amouriq@univ-nantes.fr

Florent Autrusseau
LTeN, Polytech'Nantes (U6607), and
RMeS (U1229), rue CH. Pauc,
University of Nantes, France
Florent.Autrusseau@univ-nantes.fr

***Abstract*—In order to build an intelligent dental care process that both facilitates the treatment and improves the diagnosis, an accurate tooth segmentation and recognition on panoramic X-ray images might prove helpful. Although many studies have been conducted on teeth segmentation, few methods allow to perform tooth recognition and numbering at the same time. The existing methods allowing both those processes rely on instance segmentation architectures. To fill some gaps in the area of dental image segmentation, we propose a novel approach of automatic joint teeth segmentation and numbering using the pioneer U-Net model. We are first to employ the conventional U-Net model and show its limitations to provide accurate segmentation, being affected by noisy pixels outside the teeth region and by missing teeth in the X-ray images. To overcome this problem and reduce the misclassifications, we use a bounding box prior at the level of the skip connections. Such an approach helps guiding the network to better locate the teeth, and hence improves the segmentation. To validate the effectiveness of the method, we have conducted two experiments on the DNS Panoramic Dataset: a first one using manual bounding boxes and another one relying on a preliminary step of object detection. The implemented networks were evaluated using the Dice coefficient index and our results showed that consideration of location information onto the skip connections improves the performances of the semantic segmentation by 5% to 10% in average Dice accuracy depending on the quality of the bounding box labels.**

***Index Terms*—Panoramic X-ray images, Teeth segmentation, U-Net, location prior, deep learning.**

I. Introduction

In order to formulate a diagnosis of oral diseases, panoramic X-rays are important tools for dentists to visualize the structure, shape and position of each tooth and confirm or discard a particular diagnosis such as a fracture, infection, tooth loss or simply to detect any previous dental treatment. The limitations of the image acquisition modality; its resolution, a bad contrast balance or the presence of high amplitude noise in panoramic X-rays can make the interpretation quite challenging and lead to some errors in formulating the diagnosis. So, providing an automatic analysis of panoramic X-ray images has become crucial in the dentistry community as it can increase the diagnosis accuracy while reducing the screening time and thus, the medical costs. One of such aids to dentists is medical image segmentation, which consists in the classification of each image pixel into an object of interest.

In the field of teeth segmentation, several unsupervised pixel-wise segmentation [1] approaches have been developed, mostly using intra-oral images (periapical or bitewing radiographs). Very few works have been conducted on panoramic X-ray images, including region-based [2], threshold-based [3], cluster-based [4], or boundary-based [5] methods. Silva *et al.* [1], provides an interesting overview of 10 classical methods on dental imaging segmentation. They created a dataset of 1500 panoramic X-ray images, called UFBA-UESC Dental Images Dataset, on which they tested the previously cited methods. Unfortunately, they conclude that classical segmentation solutions failed to completely isolate the teeth from neighboring bone parts present inside the mouth.

Encouraged by their recent success on various computer vision tasks, they proposed Convolutional Neural Networks (CNN) for teeth segmentation. In [1], the authors propose separating the whole dental arch from the background using deep learning. This solution surpassed most of the traditional methods considerably. Jader *et al.* [6], investigated teeth segmentation and detection on a modified version of the previously mentioned dataset using a Mask R-CNN [7] solution to carry out an instance segmentation task. However, all teeth were classified into a single category, and thus independent tooth recognition was not considered. Koch *et al.* [8] trained a semantic segmentation network using the pioneer U-Net [9], with a patching scheme strategy and obtained good results at detecting the foreground class (presence of teeth). However, again, this approach does not allow an independent tooth extraction / recognition.

Indeed, most of the recent teeth segmentation methods that can be found in the literature [10], [11] are devoted to one-class segmentation, *i.e.* all teeth are classified into one single category, thus ignoring both the morphological properties and

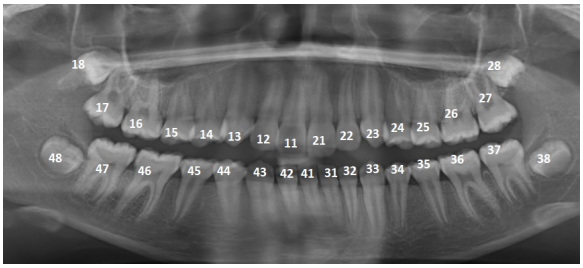


Fig. 1. FDI tooth notation

the independent tooth position. However, the teeth numbering is crucial for composing the odontogram of a patient and facilitate the diagnosis. The odontogram, or dental chart, is a diagram used by dentists to record all the needed information concerning a patient's set of teeth. On this chart, not only are the patient's teeth (or absence of teeth) represented, but also all the treatments the patient underwent. An adult commonly has 32 teeth, which are classified according to their locations (upper left, upper right, lower left and lower right quadrants of the panoramic X-ray) and positions (1 to 8 from incisors to molars). Fig. 1 shows the FDI World Dental Federation tooth notation.

It is quite challenging to come up with a computer vision strategy that will be able to precisely recognize and locate each of the 32 teeth on panoramic X-rays as such radiographs commonly suffer from a very large variability, due to a rather high amplitude noise or even low contrasts. Moreover, within each category (incisors, canines, premolars and molars), the teeth may sometimes be quite difficult to differentiate for non experts. Their shapes are very close, their grey level amplitudes are in the same range, and the position information can be misleading, as a missing tooth might lead its neighbor to shift and occupy its site. Aside from the teeth arrangement, which can vary significantly from person to person, two consecutive teeth can share some restorations and can be of very close proximity if not even overlapping on the X-ray image. This leads us to question CNN's ability to differentiate two distinct teeth sharing the same morphology and same pixel values. To automatically segment, detect and number the teeth, Silva *et al.* [12] proposed instance segmentation methods using 4 different architectures, namely, PANet, HTC, ResNeSt and Mask R-CNN, and obtained very good performances.

In this work, we propose a different approach, where by using semantic segmentation we are able to accurately segment the image and locate the teeth positions, and thus, we can efficiently number the teeth in a panoramic X-ray using a U-Net based architecture. The latter is known for its ability to improve the effect of fine-grained segmentation compared to a Fully Convolutional Network (FCN) [13] segmentation head, used in instance segmentation.

In order to increase the segmentation accuracy and reduce the misclassification of some teeth positions, we exploit the use of spatial prior onto U-Net training. For instance, recent works have studied the integration of prior knowledge, such as the shape or location of objects, used as constraints, in order to improve CNN's performance [14]–[21]. For example, Zotti *et al.* [14] proposed extracting the cardiac center of mass by adding a regression model onto the bottleneck of U-Net. In [18], the authors injected multi-scale patches extracted from the image into their network to incorporate spatial information. In this context, we explore the effect of injecting the tooth location information by introducing a bounding box prior at the level of the skip connections [21].

This paper is structured as follows: in section II, we provide some details on the modified U-Net architecture, show how it considers the teeth location for an optimized segmentation task. In section III we analyze improvements of the performances brought by this spatial localization addition onto the skip connections compared to an original U-Net model and finally, section IV concludes our work.

II. Method

In this section, we present the pipeline of our method. The architecture of the semantic segmentation model is described as well as the used prior knowledge supervision technique.

A. U-Net Architecture

As previously mentioned, the architecture of our method is based on the U-Net model, which is widely used for medical image semantic segmentation as it can achieve very good performances while relying on small sets of training data. As inputs, our model takes some images of size 512×1024 and at the output, produces a pixel-wise probability prediction for 33 classes; a separate class is dedicated to each tooth position (that is 32 classes), the last one being allocated to the background.

The U-Net architecture has a symmetric encoder/decoder structure with bottleneck and skip connections. We have chosen a 4 stages U-Net architecture. Each stage in the encoder consists of two successive convolutional blocks separated by max pooling layers in order to reduce the feature resolution, while each stage of the decoder consists of a transposed convolution followed by two convolutional blocks. Each convolutional block contains a 3×3 convolutional layer, a batch normalization layer and a Rectified Linear Unit (ReLU) activation. The last layer is a 1×1 convolution followed by a *softmax* activation function that produces the pixel-wise probability map. The encoder layers extract contextual information or features present in the image. On the other hand, the decoder layers are dedicated to determine the localization of the patterns and recover the image maps with their original input size. In order to combine both contextual and positional features, the skip connections are used between the encoder and the decoder.

Despite its success in incorporating global features via its symmetric architecture and the use of skip connections, the U-Net model has some limitations when incorporating location information, as we will see below. The way to address this issue is through the integration of prior information and more specifically, location prior, into the model architecture. In this

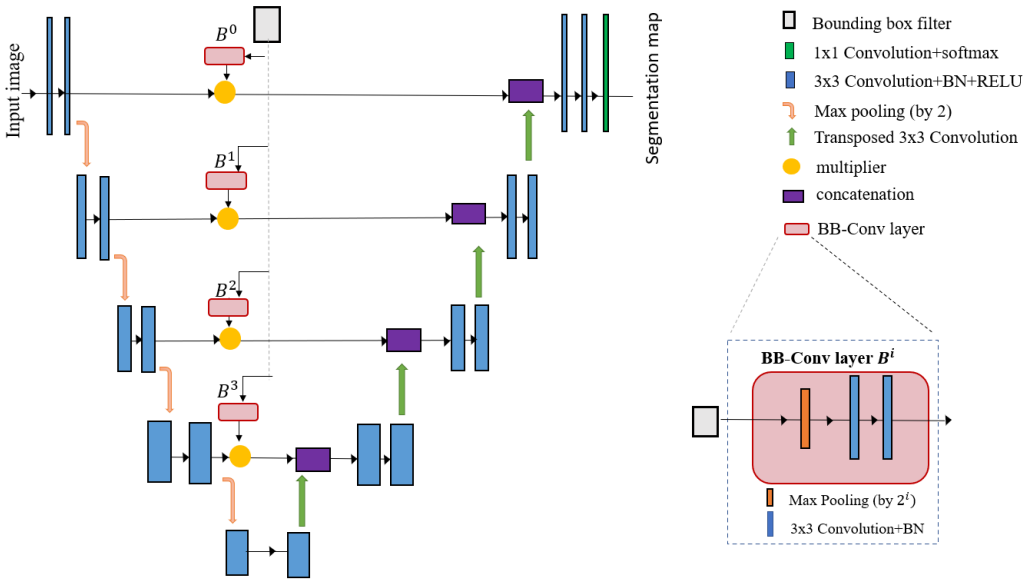


Fig. 2. Network architecture.

work, we adopt integrated location prior information into the learning process via bounding boxes at the level of the skip-connections, as introduced by [21].

B. Modified U-Net architecture

In this work, we adopt a very recent U-Net adaptation, the BB-Conv layer introduced by Rosana *et al.* [21]. A BB-Conv layer is composed of 2D max pooling layers followed by 2 convolutional layers. Bounding boxes relative to different teeth are fed independently through a multi-channel binary map onto the BB-Conv layer that gives, as an output, a feature map helping the network to improve the localization of a given tooth. The BB-conv layers are introduced onto each skip connection as shown in Fig. 2. At each stage, an element-wise multiplication is performed between the output of the BB-conv layer and the input features from the encoder before being concatenated with the features issued by the deconvolutional layers of the decoder. To distinguish this new network architecture from the original U-Net, in the following, it will be referred as the *Modified U-Net*.

C. Generating bounding boxes

In order to assess the effectiveness of the proposed model, two techniques are used to generate the bounding boxes. In the first case, the bounding boxes for each tooth were generated from the ground-truth as the smallest bounding box framing the tooth's mask. As we used manual bounding boxes for both our training and inference tasks, this setup supposedly provides the best results that can be reached, in the following it will be denoted as *Optimal U-Net*. Evidently, this approach only intends to prove the feasibility of the method and provide a glimpse of the best segmentation results one can reach. It isn't feasible in real world applications, as of course, the manual segmentation is not available. Ideally, automatic bounding boxes should be generated using an object detection framework for the training and inference steps. Our proposed system here is based on a region proposal approach, namely, the state-of-the-art Faster R-CNN model [22]. A Faster R-CNN is a single unified network consisting of two stages: the Region Proposal Network (RPN) which proposes regions of interest (ROIs), in which an object can be localized, and the ROI classifier that identifies which category an object belongs to and refines the bounding box generated by the RPN. As an extension of the Faster R-CNN, the Mask R-CNN introduces a parallel branch (or the mask branch) to the network in order to carry out instance segmentation. For the sake of comparison between segmentation masks, we propose directly using bounding box coordinates predicted by the classification branch of the Mask R-CNN network to be fed onto the skip connections of our Modified U-Net, rather than training a separate Faster R-CNN network.

D. Pipeline of the Proposed method

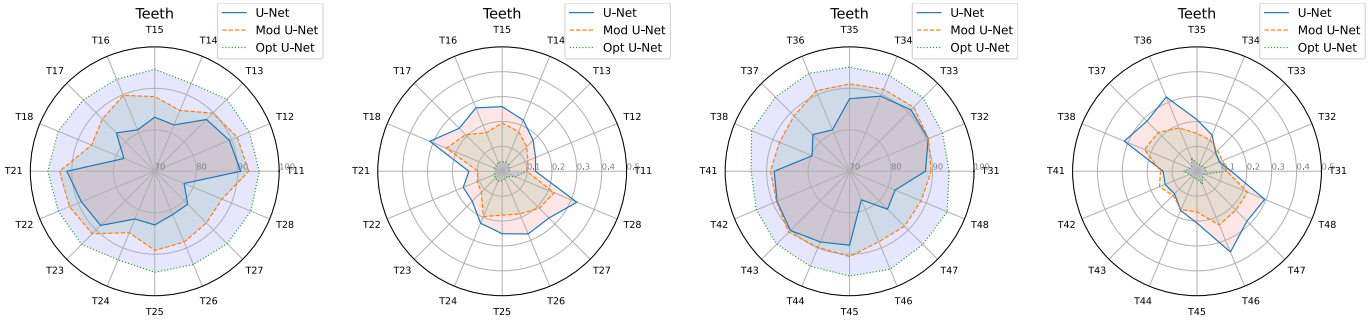
We have developed a two-steps process for teeth segmentation using a Modified U-Net architecture. In the first step, we train a Mask R-CNN model to locate the teeth, and more precisely, to extract bounding boxes of each positional tooth. In the second step, we train a Modified U-Net network using the predicted bounding boxes from the first step.

III. Experimental results

Let us now give some details on the dataset composition, provide some information on the training step and show the performances our approach can achieve.

A. Dataset

The DNS Panoramic images from IvisionLAB [12] were used. This dataset contains 543 images of size 1127×1191



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

Fig. 3. Dice coefficients (a) and (c) and Standard Deviations (b) and (d) for U-Net, Modified U-Net, and Optimal U-Net configurations ((a)&(b): teeth from the upper jaw, (c)&(d): lower jaw).

annotated with the tooth numbering (FDI notation, *i.e.* position label) using the Coco format. The images can be divided into 8 categories regarding the presence or absence of all teeth in images, of restorations and appliances (please refer to [12] for a detailed description of different categories).

Mask R-CNN and U-Net models were trained using a 4-fold cross validation. In total, 111 images were retained to build the test set and the rest of the images were divided into 4 folds (108 images each) thus composing the train and validation data in a cross-validation fashion.

B. Object detection training procedure

Our experiments are based on the open source implementation of Mask R-CNN [23]. We also benefit from transfer learning using pre-trained weights taken from the MS COCO dataset. The network's backbone was ResNet-101. We used the stochastic gradient descent optimizer with a learning rate of 0.001 and momentum of 0.9. For each combination of train/validation data, we trained a Mask R-CNN network using a batch size of 1 for 80 epochs (the validation loss didn't decrease after 80 epochs for any of the 4 networks). For training and inference, we set a detection score threshold at 0.05 to filter out all predictions with low confidence. Additional details of the several hyperparameters and parameters related to the implementation can be found in [23]. The performances of the object detection task were evaluated using the mean Average Precision (mAP), i.e. the average of the Average Precisions calculated for all tooth classes with the traditional IoU (Intersection over Union) of 0.5 [24]. The mAPs of 4 trained neural networks (of 4-CV) on their respective validation data are 0.983, 0.963, 0.980 and 0.970 respectively. These results are comparable to the performances reached by previous works on teeth detection and numbering [12], [25], [26] and show that the object detection network performs very well, and most importantly, is consistent across the validation dataset.

After obtaining all of the bounding boxes and before training our Modified U-Net, for every image in our dataset, we choose one box to be selected per class at most, i.e., the box presenting the highest confidence score, based on the natural assumption that each tooth can occur in the image only once.

C. Loss function and model training

We use Dice loss as the criterion to optimize the U-Net model parameters. This loss is widely used in medical image segmentation and is defined as :

$$L_{\text{dice}} = -\frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i=1}^{N} p_c(i).g_c(i)}{\sum_{i=1}^{N} p_c(i)^2 + \sum_{i=1}^{N} g_c(i)^2}$$

where N is the number of pixels in the image, C represents the number of class labels, $p_c(i)$ is the predicted probability of class c at pixel i and $g_c(i) \in \{0, 1\}$ represent the ground-truth label at position i. Note that $C = 33$ as we take into account the background class in the computation of the Dice loss; The idea is to improve the ability of the model to detect the boundaries of each tooth.

In order to study the effectiveness of the proposed method, we conducted three different experiments using three different model settings : a classical U-Net, a Modified U-Net with bounding boxes generated by Mask R-CNN and an Optimal U-Net described earlier. When building our dataset, we resized the images to 512×1024 pixels. For all three U-Net configurations, we used 64 feature maps in the highest stage. The number of feature maps double with each downsampling step until it reaches 512 within the fourth stage and within the bottleneck. Moreover, the Adam optimizer is used to train our model with an initial learning rate of 0.0001. We set the number of training epochs to be 60 with a batch size of 2. The learning rate was halved each 15 epochs if the validation performance did not improve.

D. Quantitative Results

Fig. 3 shows the overall results of the proposed method in terms of average Dice coefficient index (%) and its standard deviation for each tooth position from the cross validation step.

We observe that the Modified U-Net configuration consistently outperforms the original U-Net model for all teeth classes. Evidently, the Optimal U-Net model offers even better performances. The Optimal U-Net surpasses the other two

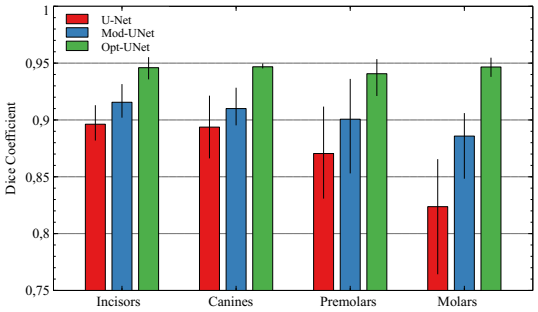


Fig. 4. DICE coefficients for the three tested CNN architectures and for the four distinct teeth types. Results obtained from the test set (composed of 111 images).

models with an average dice index of 94.5% and with minimal standard deviations. This confirms the importance of properly integrating location knowledge onto the model structure. The performance gap between the Optimal and the Modified U-Net is due to the inaccuracies between the ground-truth bounding boxes and the ones detected by the Mask R-CNN network.

The quite large Dice score difference between the original U-Net and the Modified U-Net comes from the detection inaccuracies of the molar teeth. The inclusion of prior information via the bounding boxes onto the skip connections increases the dice accuracy by over 11% on tooth #46, 10% on teeth #28 and #36, 9% on tooth #16 and by 8% on teeth #18 and #38. This stems from the fact that the original U-Net model suffers from a misclassification issue on panoramic X-ray images with noisy pixels, degraded teeth and most importantly, missing teeth. The positional information carried by the bounding boxes guarantees that all teeth are located nearby their correct location. Finally, the Modified U-Net and U-Net architectures achieve an average Dice accuracy of 89.5% and 85% respectively.

Following the same protocol, we computed the Dice metric using the test data set. Using unseen data, the Optimal U-Net offered the best performances with an average dice score of 94.49% followed by the Modified U-Net with an average Dice of 90%. Unsurprisingly, the U-Net model exhibits the worst Dice coefficient with an average of 86%. Fig. 4 summarizes the average and the boundaries (error bars) of the Dice ratio for each architecture with respect to the tooth type. As it can be observed from this Figure, the Original U-Net performances significantly decrease from incisors (0.89) to molars (0.82). This is due to teeth being segmented wrongly as one of their (missing) neighbors, which mostly occurs for premolars and molars. It is worth noting that 2.7% of incisors, 2.9% of canines, 3.8% of premolars and 7.9% of molars were missing in our test set (composed of 111 X-rays). The Modified U-Net presents a much smaller decrease in Dice ratio (0.91 for incisors, down to 0.88 for molars) explained by the rare miss-labeling of teeth by the object detection model. This can be clearly observed from Table I and Fig. 5. Among the 36 detected miss-classified teeth, 29 occurred nearby

TABLE I
RESULTS OF THE OBJECT DETECTION TASK (MASK R-CNN) ON THE TEST SET USING THE BEST NETWORK SOLUTION ACCORDING TO MAP. TOOTH DETECTION AND NUMBERING IS DONE USING BOUNDING BOXES AND AN IOU THRESHOLD OF 0.5

	Count
Total Number of teeth	***3382***
Detected and correctly classified	***3333***
Miss-classified detections	***36***
Missed detections	***13***

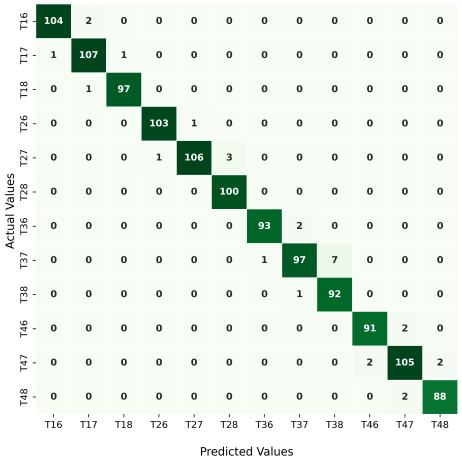


Fig. 5. Sub-matrix of the confusion matrix for a 0.5 IoU detection threshold on the test dataset, corresponding to molar teeth.

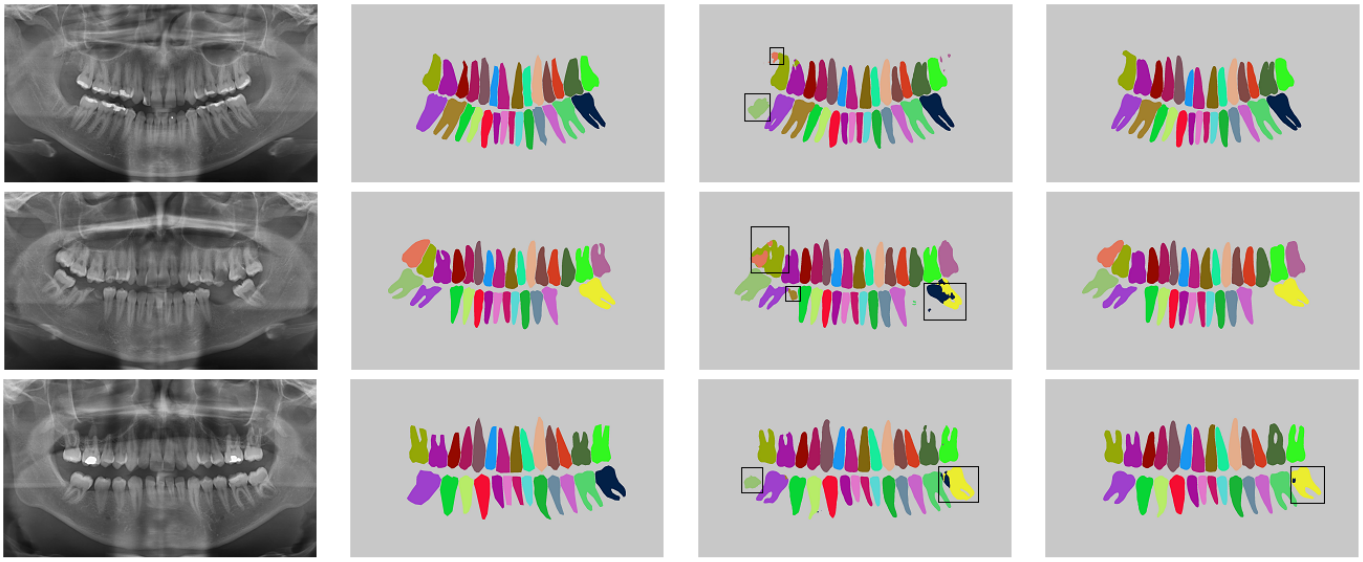
molar teeth. However, the Optimal U-Net maintains consistent performances (0.94) across the different tooth types. These results confirm, once again, the importance of integrating some geographic location information into the segmentation task.

E. Qualitative results

In Fig. 6, we show some typical segmentation results produced by both the U-Net and our proposed Modified U-Net. Compared to U-Net, the Modified U-Net avoids segmenting surrounding tissues as teeth (the first row), most importantly, it avoids classifying pixels from one class (one tooth) into 2 or more classes (the second row). This mainly occurs when a tooth is degraded or missing. The third row shows an example of a false segmentation also occurring with the Modified U-Net due to a mislabeling of tooth position by the object detection network (tooth number 37 being labeled as number 38).

F. Comparison with instance segmentation

U-Net (or some of its derived forms) is among the gold standards for biomedical image segmentation, as it presents a high robustness despite a noisy image context. However, several works propose instance segmentation using the Mask R-CNN architecture on medical images [27]–[29] as the latter offers interesting performances for some problems that require a sense of a whole object segmentation rather than focusing



(a) Original scan (b) Ground Truth (c) U-Net (d) Modified U-Net

Fig. 6. Examples of the segmentation results using the U-Net (c) and Modified U-Net (d) architectures.

on individual pixels only. In fact, R-CNN (Region Based Convolutional Neural Networks) were initially developed for object detection purposes. The Mask R-CNN extends the Faster R-CNN by adding a branch, the purpose of which, is to predict the object mask, hence providing segmentation capabilities. However, at the pixel-level, the segmentation task of Mask R-CNN is performed by FCN, which is less accurate than U-Net especially for small datasets. To address this issue, the Modified U-Net combines the performance of U-Net with the advantage of an object detection step. For a thorough evaluation of our method and an exhaustive comparison with state-of-the-art approaches, we hereby compare our results with the segmentation masks obtained by the mask branch of Mask R-CNN. Table II shows that the Modified U-Net provides higher performance in terms of dice coefficient (2 percentage point) than Mask R-CNN for all teeth types.

Finally, it is important to note that the Modified U-Net can be used easily with faster object detection approaches than region proposed methods such as the one stage detectors [30]–[32]. This method has the potential for practical and faster application for automated dental analysis, but further investigation is needed to assess its precision. On the other hand, one stage detectors are not quite suitable for instance segmentation as the mask branch incorporation is not straightforward.

TABLE II
PERFORMANCE COMPARISON OF OUR METHOD WITH MASK R-CNN ON THE TEST SET

Model	**Avg Dice(%) per class**			
Network	***Incisors***	***Canines***	***Premolars***	***Molars***
Mod U-Net	**91.55**	**91.00**	**90.00**	**88.58**
Mask RCNN	89.56	89.45	88.70	87.55

IV. CONCLUSION

The vast majority of the research works on teeth detection and segmentation focuses on periapical or intraoral radiographs. Few works are devoted to the panoramic X-rays, as such radiographs unfortunately often suffer from a poor quality (high noise, unbalanced contrasts, etc.). The segmentation of panoramic X-rays is very challenging. In this work, we propose an automatic teeth segmentation and numbering using the Modified U-Net model that integrates location prior by connecting bounding boxes to the skip connections. We have demonstrated effectively that it is crucial to incorporate some geometric localization information into the U-Net, in order to improve its performances and avoid false segmentation occurring on panoramic images with noisy surrounding pixels, degraded or missing teeth. Experiments showed an increase of the Dice coefficient index for all teeth classes with an average increase of 5% and which can go up to 10% for an Optimal U-Net. Furthermore, the proposed segmentation scheme is superior to the Mask R-CNN segmentation in terms of the Dice coefficient. As the Modified U-Net depends on bounding boxes generated by an object detection approach, the future steps include using or developing an optimized object detection algorithm that minimizes the shift between the detected *vs.* ground-truth bounding boxes.

ACKNOWLEDGMENT

This work was partially supported by the I-Site NExT fund (call for projects “Innovez”), the “Région des Pays de la Loire”, as well as the “Fonds Européen de Développement Régional” (FEDER).

REFERENCES

[1] G. Silva, L. Oliveira, and M. Pithon, "Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives," Expert Systems with Applications, vol. 107, pp. 15 – 31, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418302252

[2] A. Lurie, G. M. Tosoni, J. Tsimikas, and W. Fitz, "Recursive hierarchic segmentation analysis of bone mineral density changes on digital panoramic images," Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, vol. 113(4), pp. 549–558, 2012.

[3] Y. Y. Amer and M. J. Aqel, "An efficient segmentation algorithm for panoramic dental images," Procedia Computer Science, vol. 65, pp. 718–725, 2015, international Conference on Communications, management, and Information technology (ICCMIT'2015). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091502846X

[4] M. K. Alsmadi, "A hybrid fuzzy c-means and neutrosophic for jaw lesions segmentation," Ain Shams Engineering Journal, vol. 9, no. 4, pp. 697–706, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2090447916300430

[5] M. R. M. Razali, N. S. Ahmad, R. Hassan, Z. M. Zaki, and W. Ismail, "Sobel and canny edges segmentations for the dental age assessment," in 2014 International Conference on Computer Assisted System in Health, 2014, pp. 62–66.

[6] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic x-ray images," in 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018, pp. 400–407.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[8] T. L. Koch, M. Perslev, C. Igel, and S. S. Brandt, "Accurate segmentation of dental panoramic radiographs with U-Nets," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 15–19.

[9] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on arXiv:1505.04597 [cs.CV]). [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a

[10] Y. Zhao, P. Li, C. Gao, Y. Liu, Q. Chen, F. Yang, and D. Meng, "Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network," Knowl. Based Syst., vol. 206, p. 106338, 2020.

[11] Q. Chen, Y. Zhao, Y. Liu, Y. Sun, C. Yang, P. Li, L. Zhang, and C. Gao, "Mslpnet: multi-scale location perception network for dental panoramic x-ray image segmentation," Neural Comput. Appl., vol. 33, pp. 10 277–10 291, 2021.

[12] B. Silva, L. Pinheiro, L. Oliveira, and M. Pithon, "A study on tooth segmentation and numbering using end-to-end deep neural networks," in 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2020, pp. 164–171.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440, 2015.

[14] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P.-M. Jodoin, "Gridnet with automatic shape prior registration for automatic mri cardiac segmentation," in International workshop on statistical atlases and computational models of the heart. Springer, 2017, pp. 73–81.

[15] C. Zotti, Z. Luo, A. Lalande, and P.-M. Jodoin, "Convolutional neural network with shape prior applied to cardiac mri segmentation," IEEE journal of biomedical and health informatics, vol. 23, no. 3, pp. 1119–1128, 2018.

[16] R. Trullo, C. Petitjean, D. Nie, D. Shen, and S. Ruan, "Joint segmentation of multiple thoracic organs in ct images with two collaborative deep architectures," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2017, pp. 21–29.

[17] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan et al., "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," IEEE transactions on medical imaging, vol. 37, no. 2, pp. 384–395, 2017.

[18] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. van Uden, C. I. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel, "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," Scientific Reports, vol. 7, no. 1, pp. 1–12, 2017.

[19] H. Ravishankar, S. Thiruvenkadam, R. Venkataramani, and V. Vaidya, "Joint deep learning of foreground, background and shape for robust contextual segmentation," in International Conference on Information Processing in Medical Imaging. Springer, 2017, pp. 622–632.

[20] H. Oda, H. R. Roth, K. Chiba, J. Sokolić, T. Kitasaka, M. Oda, A. Hinoki, H. Uchida, J. A. Schnabel, and K. Mori, "Besnet: boundary-enhanced segmentation of cells in histopathological images," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 228–236.

[21] R. El Jurdi, C. Petitjean, P. Honeine, and F. Abdallah, "BB-UNet: U-Net with bounding box prior," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 6, pp. 1189–1198, 2020.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pp. 1137–1149, 2016.

[23] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on keras and tensorflow," 2017.

[24] M. Everingham and L. Van Gool, "L. and williams," CKI and Winn, J. and Zisserman, A., The PASCAL Visual Object Classes Challenge, 2008.

[25] D. V. Tuzoff, L. N. Tuzova, M. M. Bornstein, A. S. Krasnov, M. A. Kharchenko, S. I. Nikolenko, M. M. Sveshnikov, and G. B. Bednenko, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," Dento maxillo facial radiology, vol. 48 4, p. 20180051, 2019.

[26] F. Mahdi, K. Motoki, and S. Kobashi, "Optimization technique combined with deep learning method for teeth recognition in dental panoramic radiographs," Scientific Reports, vol. 10, 11 2020.

[27] R. Anantharaman, M. Velazquez, and Y. Lee, "Utilizing mask r-cnn for detection and segmentation of oral diseases," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2197–2204.

[28] J.-H. Shu, F.-D. Nian, M.-H. Yu, and X. Li, "An improved mask r-cnn model for multiorgan segmentation," Mathematical Problems in Engineering, vol. 2020, p. 8351725, 2020. [Online]. Available: https://doi.org/10.1155/2020/8351725

[29] R. O. Dogan, H. Dogan, C. Bayrak, and T. Kayikcioglu, "A two-phase approach using mask r-cnn and 3d u-net for high-accuracy automatic segmentation of pancreas in ct imaging," Computer Methods and Programs in Biomedicine, vol. 207, p. 106141, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721002169

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.