A study on tooth segmentation and numbering using end-to-end deep neural networks

Bernardo Silva, Laís Pinheiro, Luciano Oliveira
Intelligent Vision Research Lab,
Universidade Federal da Bahia, Salvador, Bahia
Email: http://ivisionlab.ufba.br/people/

Matheus Pithon
Universidade Estadual do Sudoeste da Bahia, Jequié, Bahia
Email: matheuspithon@gmail.com

***Abstract*—Shape, number, and position of teeth are the main targets of a dentist when screening for patient's problems on X-rays. Rather than solely relying on the trained eyes of the dentists, computational tools have been proposed to aid specialists as decision supporter for better diagnoses. When applied to X-rays, these tools are specially grounded on object segmentation and detection. In fact, the very first goal of segmenting and detecting the teeth in the images is to facilitate other automatic methods in further processing steps. Although researches over tooth segmentation and detection are not recent, the application of deep learning techniques in the field is new and has not reached maturity yet. To fill some gaps in the area of dental image analysis, we bring a thorough study on tooth segmentation and numbering on panoramic X-ray images by means of end-to-end deep neural networks. For that, we analyze the performance of four network architectures, namely, Mask R-CNN, PANet, HTC, and ResNeSt, over a challenging data set. The choice of these networks was made upon their high performance over other data sets for instance segmentation and detection. To the best of our knowledge, this is the first study on instance segmentation, detection, and numbering of teeth on panoramic dental X-rays. We found that (i) it is completely feasible to detect, to segment, and to number teeth by through any of the analyzed architectures, (ii) performance can be significantly boosted with the proper choice of neural network architecture, and (iii) the PANet had the best results on our evaluations with an mAP of 71.3% on segmentation and 74.0% on numbering, raising 4.9 and 3.5 percentage points the results obtained with Mask R-CNN.**

***Index Terms*—deep neural networks, instance segmentation and numbering, panoramic dental X-rays**

I. Introduction

X-rays are a valuable resource that helps dentists diagnose cysts, tumors, fractures, and other conditions that require more information than that is possible to gather by directly examining the patient [1]. However, interpreting an X-ray is not a trivial task, usually taking years of training and experience before a professional can give reliable reports. It is a time-consuming and prone to errors task that requires many aptitudes from the professional. Another potential issue is that there are many different types of X-rays, each having peculiarities and challenges that make the screening difficult [2].

Due to the aforementioned reasons, some methods have been proposed to perform automatic analyses over X-rays using **tooth detection and classification** [3, 4, 5, 6], **tooth segmentation** [7, 8, 9, 10], **caries detection** [11, 12, 13, 14], **osteoporosis diagnosis** [15, 16, 17], and **forensic identification** [18, 19, 20]. These methods usually intend to carry out analyses on the most common types of dental X-rays: periapical, bitewing, and panoramic.

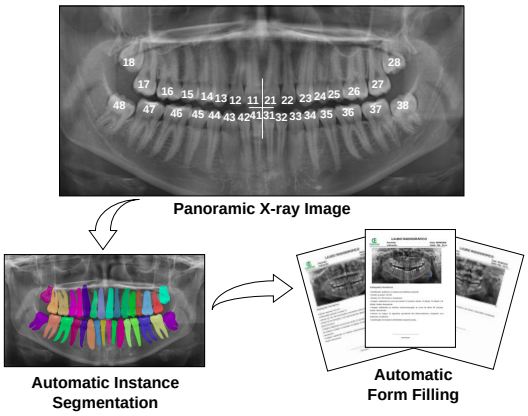


Fig. 1. The FDI (French acronym for World Dental Federation) two-digit notation is used by dentists to report their findings, where number, shape, and location of teeth are essential information to support decisions in dentistry.

According to Silva et al. [8], before 2018, most academic research on automatic analysis of dental X-rays based their studies on handcrafted feature extractors. The majority of these works also overlooked panoramic X-rays, probably due to the challenging nature of this type of image. Indeed, differently from periapical and bitewing X-rays, panoramic images include several parts of the body like teeth, gums, jaw, skull, spine, and other bones. These characteristics, along with the low contrasts, turn handcrafted strategies unsuitable for panoramic images.

Silva et al. [8] also examined several handcrafted feature extractors for semantic segmentation of teeth and concluded that these methods perform poorly compared to a state-of-the-art deep learning network. Under these circumstances, where the deep learning techniques offer the most promising avenue to medical image analysis, we investigate end-to-end neural networks' feasibility for tooth segmentation and numbering on panoramic X-rays. This is a rather relevant task since shape, position, and type of teeth are the main targets of dentists and many computer methods for X-ray analysis. For instance, the

2377-5416/20/$31.00 ©2020 IEEE
DOI 10.1109/SIBGRAPI51738.2020.00030

first findings in dental reports are missing teeth, which are usually specified in a two-digit notation (see Fig. 1).

A. Related works

The advent of deep learning techniques boosted the performance of automatic image analysis methods. This is rather remarkable on the panoramic dental X-rays due to the traditional methods' poor performance on such challenging images. In the following, we detail the studies on the matter, according to the image task.

Semantic segmentation: Silva et al. [8] present a thorough literature review on classic methods to segment teeth on panoramic dental X-rays. They also introduce a considerably large and diverse data set, called UFBA-UESC Dental Images data set. Experiments on this data set allowed the authors to conclude that the Mask R-CNN solution [21], in which the whole dental arch is considered a single instance, is by far better than classic methods. That was the first study to demonstrate the advantages of applying deep learning on panoramic X-rays. Koch et al. [10] trained a U-Net [22] on the UFBA-UESC Dental Images data set, where horizontal flipping and model ensemble improved performance. Both solutions surpassed the results of classic methods. However, semantic segmentation does not provide the necessary details for further processing steps in most of the automatic dental analysis.

Detection and segmentation: Jader et al. [9] were the first to investigate detection and segmentation of teeth on panoramic X-rays, although without numbering. They modified the UFBA-UESC Dental Images data set to include information of tooth instances, coining this new data set as UFBA-UESC Dental Images Deep. This was done by manually separating the teeth of the binary masks that represented the entire dental arch. A Mask R-CNN, backboned by a ResNet-101, was trained and validated with 193 and 83 images, respectively. The final network was then tested in the remaining 1224 images of the data set in a binary pixel-wise fashion, and an F_1 score of 88% was reported.

Detection and numbering: Tuzoff et al. [4] were the first to apply deep learning to detect and number teeth on panoramic X-rays. This method is different from the approach proposed by Jader et al. [9], since the latter detects and segment the teeth, but does not number them. The work of Tuzoff et al. [4] used two neural networks. The first one, a Faster R-CNN [23], outputs bounding boxes. Then, regions of the input image are cropped accordingly, feeding the second network – a VGG-16 [24] –, which numbers each tooth into one of the 32 possibilities. Finally, heuristics guarantee the consistency of the numbering. Tuzoff et al. [4] evaluated this two-stage approach on a data set containing 1352 images for training and 222 images for testing. This complex solution does not allow for end-to-end training, and heuristics could drive the results to undesirable traits. Under this perspective, the method proposed by Chung et al. [6] brings some advantages. The method consists of a neural network based on point-wise tooth localization that first performs center point regression of 32 points. Each point represents the location of a single tooth in an adult mouth, regardless of its presence, which allows for the automatic assignment of a label. Later, the tooth centers are refined, and bounding boxes are determined in a cascade way. However, the proposed method was trained on a data set smaller than the one used by Tuzoff et al. [4].

B. Contributions

Table I summarizes related works according to the number of training images (**#Train images**), the number of validation images (**#Validation images**), the number of testing images (**#Test images**), the deep learning method used (**DL method**), and if the work proposes to detect (**Detection**), to segment (**Segmentation**) and/or to number (**Numbering**) teeth on panoramic X-rays. It is worth noting that no prior work simultaneously investigated tooth segmentation, detection, and numbering. Bare all that in mind, we propose to investigate the performance of several end-to-end deep learning networks, specialized in instance segmentation, on panoramic X-rays. This is done through a benchmark that includes four network architectures chosen for having reached state-of-the-art results on the COCO data set [25]. They are: Mask R-CNN [21], Path Aggregation Network [26], Hybrid Task Cascade [27], and Split Attention Network [28]. The performance of these networks was assessed over a modified version of UFBA-UESC Dental Images Deep data set [9], specially annotated to have segmentation and numbering information. To the best of our knowledge, our work is the first one to exploit instance segmentation with numbering on a considerably large panoramic X-ray data set.

II. Segmenting and numbering teeth by end-to-end deep networks

A healthy adult mouth contains 32 teeth, which are designated by their functions (incisors, canines, premolars, and molars), quadrants (upper right, upper left, lower left, lower right), and positions (first, second, third). Because the teeth names are so long, it turned necessary to use a concise notation to expedite the filling of forms, being the FDI World Dental Federation notation the most common one (see Fig. 1). The FDI notation specifies the permanent teeth through a two-digit number: the first digit indicates a quadrant, while the second one represents the tooth type. Dentists use this notation, along with the teeth location and shape, to make their analyses and report their findings. These analyses can be supported and automated by instance segmentation neural networks.

Today's state-of-the-art methods for instance segmentation are heavily based on object detection techniques, which are usually grouped into two families: the one-stage detectors and the two-stage detectors. As the name suggests, the one-stage detectors can localize and classify all objects in a single stage, having the YOLO network [29] as the most distinguishable member of the family. On the other hand, the two-stage detectors, mainly represented by the Faster R-CNN [23], propose regions of interest (RoIs) in their first stage. In

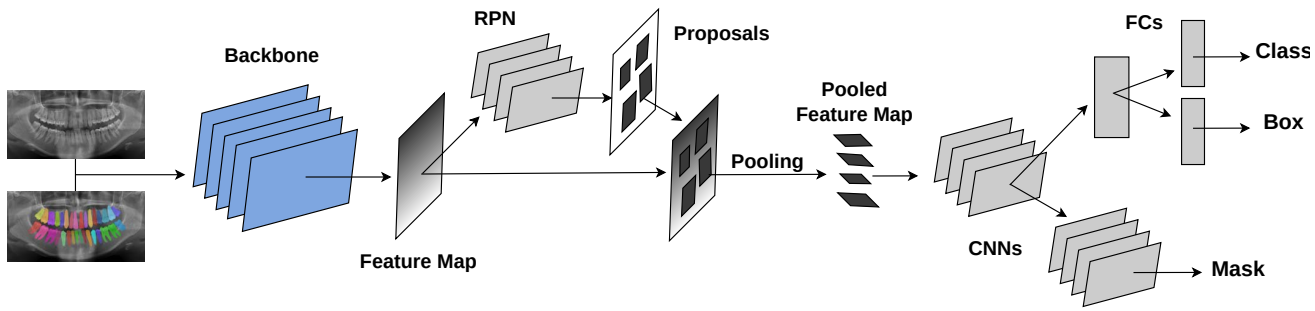


Fig. 2. General structure of an end-to-end instance segmentation network based on a two-stage detector.

TABLE I
SUMMARY OF WORKS ON SEGMENTATION, DETECTION AND NUMBERING ON PANORAMIC DENTAL X-RAY IMAGES.

Paper	#Train images	#Validation images	#Test images	DL method	Detection	Segmentation	Numbering
Silva et al. [8]	753	452	295	Mask R-CNN		✓	
Koch et al. [10]	1200	-	300	U-Net		✓	
Jader et al. [9]	193	83	1224	Mask R-CNN	✓	✓	
Tuzoff et al. [4]	1352	-	222	Faster and VGG	✓		✓
Chung et al. [6]	574	162	82	Point-wise Localization	✓		✓
Ours	324	108	111/778	Several	✓	✓	✓

the second stage, these regions are labeled, and their bounding boxes are refined.

Although slower than the one-stage detectors (which can be even employed in on-the-fly applications), the two-stage detectors are usually more precise, and the incorporation of a mask prediction branch in their structure is straightforward. These characteristics make the two-stage detectors a more suitable choice for instance segmentation on panoramic X-rays because this type of application requires high precision, but not necessarily speed. Hence, we only considered the two-stage networks, which usually rank on the top of benchmarks.

The general structure of a two-stage instance segmentation network is depicted in Fig. 2. A backbone extracts features, which are then passed to a region proposal network (RPN). Through the use of anchors, the RPN determines RoIs, in which an object can be localized. For each RoI, there is a corresponding volume of the feature map that needs to be converted into a fixed shape through some pooling technique. These fixed shape volumes feed the second stage of the network, which uses fully connected layers to predict the classes and refine the bounding boxes. In the case of instance segmentation, an additional branch, usually comprised of convolutional layers, is used for mask prediction.

To build our benchmark, we selected neural networks that reached state-of-the-art performance by introducing new techniques for the components of the two-stage detectors (backbone, pooling technique, mask predictor). In order to conduct a fair comparative analysis, we fixed the backbone to be the ResNet-50 [30], because all chosen architectures have available implementations with it. The exception is due to the ResNeSt architecture [28]. In this case, the work's novelty is the backbone itself, so we performed our experiments with ResNeSt-50 since it has a similar number of weights as ResNet-50. In the next sections, we briefly discuss the four chosen neural network architectures.

A. Mask R-CNN

The Mask R-CNN [21] was the first network to extend the Faster R-CNN for instance segmentation, surpassing, even in its basic implementations, the winners of the COCO 2016 Challenge on the instance segmentation task. This was attained with few improvements in the Faster R-CNN structure, namely: (i) a mask predictor branch, (ii) the incorporation of the feature pyramid network (FPN) [31] in the backbone, and (iii) the development of the RoiAlign technique.

The mask predictor branch generates low-resolution binary masks (typically 28×28) for each instance through fully convolutional layers. The FPN modifies the backbone by adding a top-down path that is created by lateral connections with the original backbone. All that allows predictions with rich semantic features and spatial information at all scales, improving object detection. He et al. [21] also introduced the RoiAlign strategy – a quantization-free pooling technique that preserves the spatial locations, while significantly impacting the mask predictions.

B. Path Aggregation Network

The Path Aggregation Network (PANet) [26] was the winner architecture of the COCO 2017 Challenge on the instance segmentation task. PANet owns a core structure different from

TABLE II
CHARACTERISTICS OF THE DATA SETS USED IN OUR STUDY.

Cat.	32 Teeth	Restoration	Appliance	Num & IS	Segm.
1	✓	✓	✓	23	57
2	✓	✓		174	80
3	✓		✓	42	11
4	✓			92	68
7		✓	✓	36	87
8		✓		128	355
9			✓	14	33
10				34	87
Total				543	778

the Mask R-CNN in few aspects: (i) the addition of a bottom-up path in the backbone, (ii) the development of the adaptive feature pooling (AFP), and (iii) the use of tiny fully connected layers in the mask prediction branch. The bottom-up path uses lateral connections with the FPN backbone aiming to ease information propagation, which can pass through more than 100 layers in some backbones. Differently from the RoiAlign, the AFP allows the combination of features of all levels of the hierarchy, letting the network to pick the best. Finally, tiny fully connected layers aim to provide spatial location notion to the mask prediction branch.

C. Hybrid Task Cascade

The winner architecture of the COCO 2018 Challenge on the instance segmentation task was the Hybrid Task Cascade (HTC) network [27]. The main novelty of the HTC architecture is a cascade framework that pursues better performance by interweaving the object detection and segmentation, rather than in parallel, as the conventional cascade solutions [32]. The information flow is also modified through direct branches between the previous and the next mask predictors. The architecture also contains a fully convolutional branch that enhances spatial context, which can improve performance by better discriminating instances from clutter backgrounds.

D. Split-Attention Network

Zhang et al. [28] propose the Split-Attention Network (ResNeSt), which, combined with a Cascade R-CNN [32], achieves today's state-of-the-art performance on instance segmentation on the COCO data set. Therefore, the performance was boosted by developing a more powerful backbone rather than designing specific segmentation techniques. The ResNeSt architecture accomplishes this by stacking ResNet-like blocks that incorporate attention mechanisms. This produces better results than the ResNet alone, especially in downstream applications, such as object detection and segmentation, since the ResNet was specifically conceived for image classification. Due to the split-attention blocks' modularity, we find different versions of ResNeSt, such as ResNeSt-50 and ResNeSt-101. Our experiments with the ResNeSt are carried out with a Mask R-CNN backboned by a ResNeSt-50, which we simply refer to as ResNeSt.

III. MATERIALS AND METHODS

After selecting the neural networks for the benchmark, a protocol was established. The rationale was to make the most of each architecture, yet keeping fair comparison conditions. In this sense, the same criteria were applied for all, with symmetrical parameters and hyperparameters for training, as well as for inference, when it was possible. Next, we detail the methodology for implementing the chosen neural networks, the data set used, and the training and evaluation protocols.

A. Neural network implementations

The codes for all networks were gathered from public repositories[1], using open-source implementations. This way, we could avoid any mistake in implementing the networks, purely focusing on the result analysis. We also benefited from transfer learning strategies, using weights provided by the authors of the networks. This was particularly helpful for our study since the used data set consists of a few hundred images in contrast to the hundred thousand ones from the COCO data set used to provide all the network weights. The training hyperparameters, mainly from the optimizer (learning rate and batch size), were all homogenized to keep the symmetry of the comparisons. The optimizer was chosen to be the stochastic gradient descent (SGD), while the momentum value was set to 0.9 to comply with a standard value usually used in the implementations. The weight decay regularization was removed because it worsened the first experiments' performance with the training and validation data sets. The base learning rate for all original implementations were 0.02 with a batch size of 16, but it was linearly adjusted to a batch size of 4 (learning rate ended up with a 0.005 value). This was done to train all neural networks in a single Titan 24 GB GPU. Also, the several parameters and hyperparameters related to Faster R-CNN were kept consistent. Fortunately, with few exceptions, standard choices were already in use by the original implementations.

B. Data sets

Until recently, only small sets of images were available for dental image analyses, and almost all of which were intraoral X-rays (bitewing or periapical). To fill this gap, Silva et al. [8] published the UFBA-UESC Dental Images data set, which turned out to be a valuable resource for the community. This data set is comprised of 1500 high-variability panoramic images grouped into ten categories. However, the data set was originally published along with annotations only for semantic segmentation, where binary masks pixel-wisely separate teeth from non-teeth. Later, Jader et al. [9] adapted the UFBA-UESC Dental Images data set to include instance segmentation information and used 276 images containing 32 teeth for train and validation, and the remaining 1224 images for testing. This data set was called UFBA-UESC Dental Images Deep.

[1]Original source codes and weights can be found at:
Mask R-CNN – https://github.com/facebookresearch/detectron2
PANet – https://github.com/ShuLiu1993/PANet
HTC – https://github.com/open-mmlab/mmdetection
ResNeSt – https://github.com/zhanghang1989/detectron2-ResNeSt

TABLE III
RESULTS ON THE VALIDATION DATA SET (4-FOLD CROSS VALIDATION WITH 108 IMAGES PER FOLD).

		Numbering			Instance segmentation		
Network	**Best Epoch (Avg.)**	**mAP**	**AP50**	**AP75**	**mAP**	**AP50**	**AP75**
Mask R-CNN	25	68.8±0.4	**98.6**±0.4	83.7±0.8	63.4±0.4	**98.4**±0.4	79.1±0.9
PANet	32	**74.2**±0.6	98.5±0.5	**89.0**±1.0	**71.9**±0.5	**98.4**±0.5	**88.0**±1.0
HTC	14	70.2±0.7	98.4±0.5	85.9±1.1	63.8±0.5	98.3±0.6	80.4±1.3
ResNeSt	40	73.3±0.4	98.2±0.4	88.5±0.3	70.3±0.8	98.0±0.4	86.8±1.2

TABLE IV
RESULTS ON THE TEST DATA SET (INSTANCE SEGMENTATION AND NUMBERING – 111 IMAGES).

		Numbering			Segmentation		
Network	**Threshold**	**mAP**	**AP50**	**AP75**	**mAP**	**AP50**	**AP75**
Mask R-CNN	0.675±0.025	70.5 ± 1.1	97.2 ± 0.2	86.9 ± 1.8	66.4 ± 0.7	96.9 ± 0.2	85.1 ± 1.0
PANet	0.725±0.025	**74.0** ±0.1	**99.7** ±0.2	**89.2** ±0.4	**71.3** ±0.3	**97.5** ±0.3	**88.0** ±0.2
HTC	0.563±0.022	71.1 ± 0.5	97.3 ± 0.1	87.6 ± 0.5	63.7 ± 1.4	97.0 ± 0.0	82.2 ± 2.0
ResNeSt	0.663±0.022	72.1 ± 0.3	96.8 ± 0.2	87.3 ± 0.5	68.9 ± 0.5	0.966 ± 0.3	85.5 ± 0.4

TABLE V
RESULTS ON THE SEMANTIC SEGMENTATION DATA SET (778 IMAGES).

Network	**Accuracy**	**Specificity**	**Precision**	**Recall**	**F1-Score**
Mask R-CNN	96.2 ± 0.1	98.6 ± 0.1	94.1 ± 0.2	86.6 ± 0.4	90.2 ± 0.2
PANet	**96.7** ± 0.1	**98.7** ± 0.2	**94.4** ± 0.6	89.1 ± 0.9	**91.6** ± 0.2
HTC	96.0 ± 0.2	98.5 ± 0.1	93.7 ± 0.4	85.9 ± 0.8	89.6 ± 0.4
ResNeSt	96.5 ± 0.1	98.3 ± 0.0	93.0 ± 0.1	**89.5** ± 0.2	91.2 ± 0.1
Mask R-CNN [9]	95.9	98.9	94.8	84.2	89.2

For our experiments, we used the UFBA-UESC Dental Images Deep data set, after making some changes. Table II summarizes the information from the data sets used regarding the presence of **32 teeth**, **Restoration** and **Appliance**, and the number of images used for **Numbering and Instance segmentation (Num & IS)** and **Semantic segmentation (Segm.)**. Since we required tooth number information, we annotated 543 images with number information (which included the 276 used by Jader et al. [9]), keeping 778 images to evaluate semantic segmentation. We discarded the images from categories 5 and 6 due to the presence of implants and deciduous teeth. We are providing our data set, upon request, under the name DNS (Detection, Numbering, and Segmentation) Panoramic Images[2].

C. Training and evaluation protocols

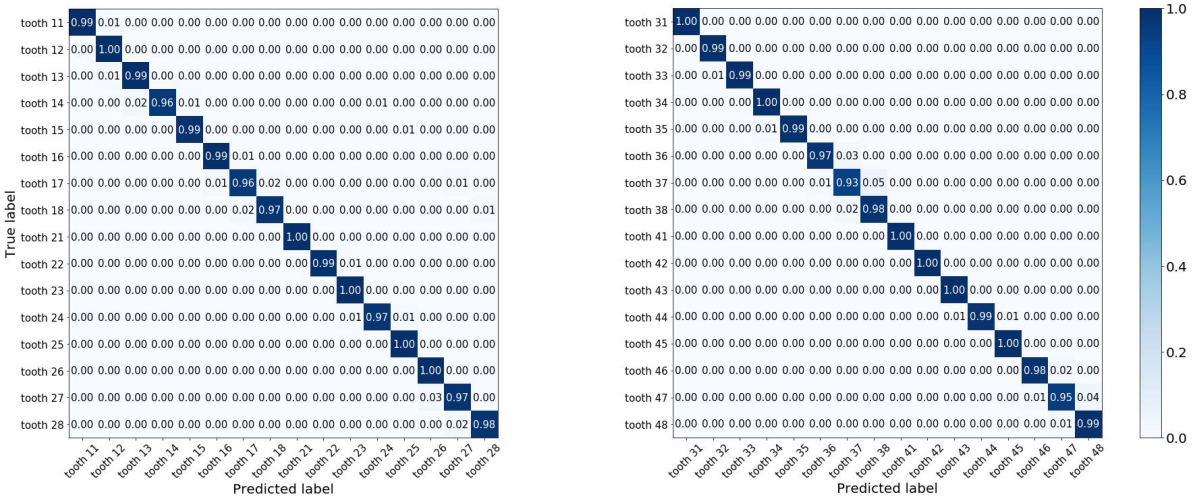
We performed two quantitative experiments. The first experiment was oriented to an instance-aware evaluation, where we used average precision with 0.5 and 0.75 cut-offs (**AP50** and **AP75**), and the mean average precision (**mAP**), with the latter being the primary metric. We applied these three metrics using the bounding boxes and masks, resulting in values for numbering and instance segmentation. The second quantitative experiment, which can be considered supplementary, was a semantic evaluation done in a pixel-wise fashion following Jader et al. [9]. F_1 score was considered the reference metric since it gives a better sense of the overall performance. To conduct a robust analysis, a cross-validation procedure was carried out. This way, we split our data into five-folds, where each fold contains approximately 20% of each category images. One of these folds was fixed as the test data set (111 images), and the other four folds (108 images in each) compose our train and validation data sets in a cross-validation fashion. For each architecture and combination, a neural network was trained. This procedure resulted in 16 neural networks (four models per architecture).

The training was conducted with no time concerns. All weights were kept unfrozen, and a linear warm-up strategy was used during the first ten epochs. The single applied augmentation technique was horizontal flipping, which was used both on training and validation, carefully changing teeth numbers to their corresponding new values (left teeth numbers turned into the right numbers and vice-versa). We proceeded with the training for 100 epochs because all models have shown clear signs of overfitting during this interval. At the end of each epoch, validation loss was computed, and the model with the smallest one was kept. Prior to the evaluation, we had to specify the detection score threshold to be used in testing. This was done by varying the threshold from 0.05 to 0.95 in steps of 0.05 and selecting the value which minimized the absolute error of the number of predicted instances and ground truth instances in the validation data set. After that, we performed all evaluations, which we discuss in the following section.

IV. EXPERIMENTS AND RESULTS

After the training procedure, we applied the instance-aware evaluation to each model on its corresponding validation

[2]Instructions on how to request the DNS Panoramic Images data set can be found at https://github.com/IvisionLab/dns-panoramic-images



(a) (b)

Fig. 3. Confusion matrices of the average results of the PANet architecture split into 2 parts: (a) superior dental arch, (b) inferior dental arch.

data set, using the training score threshold (0.05). Table III contains the summary of the results for each architecture, which includes the average epoch that resulted in the best validation loss (best epoch), and the average and standard deviation values of AP50, AP75, and mAP for numbering and instance segmentation. In this evaluation, the best performance came from PANet architecture, which won in all metrics, except in AP50 for numbering (small margin). The Mask R-CNN won with AP50 on the segmentation case, tied with PANet. However, computing AP50 for the numbering has shown to be a rather loose metric under these conditions, since all architectures reached values above 98%.

Following the evaluation protocol, we computed all instance metrics using the test data set. Table IV summarizes the mean and standard deviation of the metrics used, as well as the threshold score found for each architecture. Under these conditions (unseen data), the PANet architecture had the best performance in all metrics. It is also remarkable that the margin to the runner-up (ResNeSt) in the mAP metric was large (1.9 percentage point on numbering and 2.4 percentage point on segmentation). Figure 3 illustrates confusion matrices of the winning architecture, PANet, computed with an IoU threshold of 0.5, considering the average results for all models trained on the testing data set. We split the results into two matrices for better visualization: one for the superior dental arch and another for the inferior dental arch.

To compare our results with [9], we ran final experiments on a set of images considering annotations for semantic segmentation. Table V shows the mean and standard deviation results of each architecture following the metrics used in Jader et al. [9]. We also included the evaluation of the Mask R-CNN solution by Jader et al. [9] on the same 778 images, utilizing the weights provided by the authors.

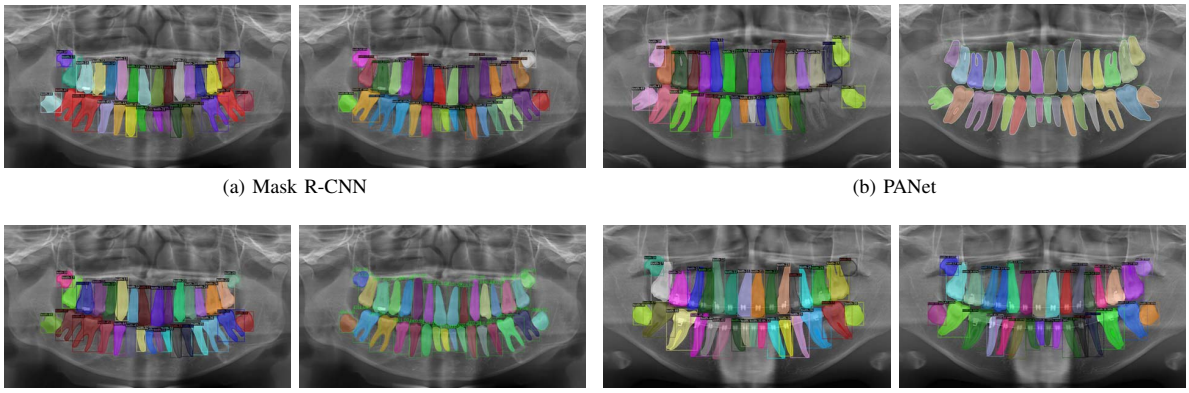
V. Discussion and Conclusions

Our results showed that instance segmentation and numbering are feasible to be accomplished by an end-to-end deep network. In our experiments, PANet achieved the best results, reaching an mAP of 74.0% on numbering, 71.3% on segmentation, and a pixel-wise F_1 score of 91.65% on semantic segmentation, all of them with minimal standard deviations. It is worth noting that PANet was also 2 percentage points better than the runner-up network on semantic segmentation. It also surpassed the Mask R-CNN architecture by 3.5 and 4.9 points on numbering and segmentation, respectively. We consider all these results remarkable since, so far, Mask R-CNN has been the most common choice for instance segmentation.

Another important aspect to be noted is that all four analyzed networks reached a better F_1 score than the Mask R-CNN model used in [9]. We must concede, however, that this is not a completely symmetric comparison since the network models here had a larger training and validation data sets with higher variability of images. On the other hand, Jader et al. [9] performed their experiments with a deeper backbone (ResNet-101) and proposed a solution that does not number the teeth, which eases the task.

Considering the results found in the confusion matrices depicted in Fig. 3, PANet correctly classified the teeth more than 90% of the times. The rare misclassifications were due to teeth being numbered as one of their neighbors, mainly occurring between the second and third molars. We also gathered the best and worst image results obtained with all the evaluated networks, considering instance segmentation and numbering (see Figs. 4 and 5). We can observe that the models perform best on healthy mouths with teeth in good conditions. The worst results come from mouths with degraded teeth, mislabeling of teeth, and crude annotations.

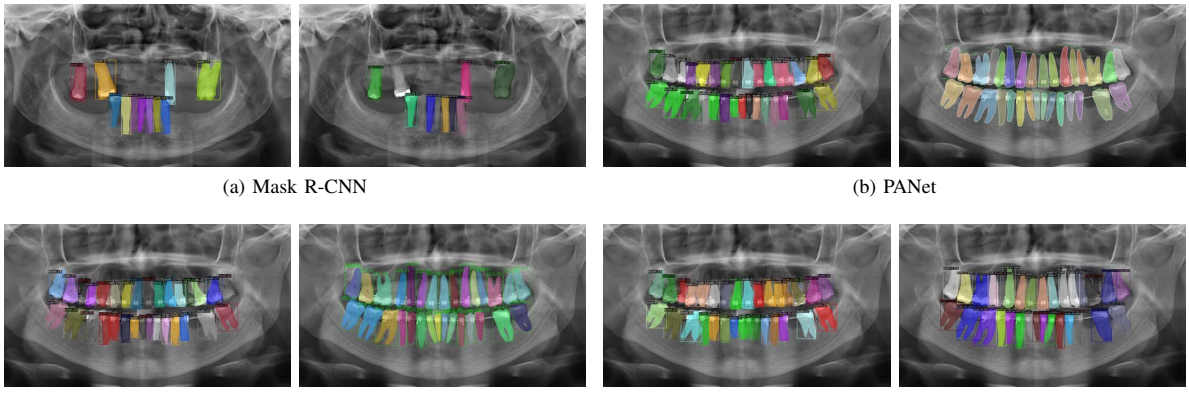
Future works should go toward the augmentation of the data



(a) Mask R-CNN (b) PANet

(c) HTC (d) ResNeSt

Fig. 4. The best result of each architecture according to mAP (ground truth is on the left side).



(a) Mask R-CNN (b) PANet

(c) HTC (d) ResNeSt

Fig. 5. The worst result of each architecture according to mAP (ground truth is on the left side).

set annotations, analyses on deciduous teeth, and improvements on the numbering performance, possibly considering the geometrical relationship between nearby teeth.

ACKNOWLEDGMENT

This work was supported by Fundação de Apoio à Pesquisa do Estado da Bahia (FAPESB), which provided scholarship to Bernardo Silva under grant BOL0569/2020. The Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) supported Luciano Oliveira under grant 307550/2018-4.

REFERENCES

[1] M. Masthoff, M. Gerwing, M. Masthoff, M. Timme, J. Kleinheinz, M. Berninger, W. Heindel, M. Wildgruber, and C. Schülke, "Dental imaging–a basic guide for the radiologist," in *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 191, no. 03. Georg Thieme Verlag KG, 2019, pp. 192–198.

[2] B. Hermanson, G. Burgdorf, J. Hatton, D. Speegle, and K. Woodmansey, "Visual fixation and scan patterns of dentists viewing dental periapical radiographs: an eye tracking pilot study," *Journal of endodontics*, vol. 44, no. 5, pp. 722–727, 2018.

[3] P. Lin, Y. Lai, and P. Huang, "An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information," *Pattern Recognition*, vol. 43, no. 4, pp. 1380–1392, 2010.

[4] D. Tuzoff, L. Tuzova, M. Bornstein, A. Krasnov, M. Kharchenko, S. Nikolenko, M. Sveshnikov, and G. Bednenko, "Tooth detection and numbering in panoramic radiographs using convolutional neural networks," *Dentomaxillofacial Radiology*, vol. 48, no. 4, 2019.

[5] H. Chen, K. Zhang, P. Lyu, H. Li, L. Zhang, J. Wu, and C. Lee, "A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[6] M. Chung, J. Lee, S. Park, M. Lee, E. Lee, J. Lee, and G. Shin, "Individual tooth detection and identification from dental panoramic x-ray images via point-wise localization and distance regularization," *arXiv preprint*

arXiv:2004.05543, 2020.

[7] Y. Amer and M. Aqel, "An efficient segmentation algorithm for panoramic dental images," *Procedia Computer Science*, vol. 65, pp. 718–725, 2015.

[8] 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.

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

[10] T. Koch, M. Perslev, C. Igel, and S. Brandt, "Accurate segmentation of dental panoramic radiographs with u-nets," in *International Symposium on Biomedical Imaging*. IEEE, 2019, pp. 15–19.

[11] A. Wenzel, "Computer-automated caries detection in digital bitewings: consistency of a program and its influence on observer agreement," *Caries research*, vol. 35, no. 1, pp. 12–20, 2001.

[12] J. Oliveira and H. Proença, "Caries detection in panoramic dental x-ray images," in *Computational Vision and Medical Image Processing*, 2011, pp. 175–190.

[13] N. Karimian, H. Salehi, M. Mahdian, H. Alnajjar, and A. Tadinada, "Deep learning classifier with optical coherence tomography images for early dental caries detection," in *Lasers in Dentistry*, vol. 10473, 2018.

[14] J. Lee, D. Kim, S. Jeong, and S. Choi, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm," *Journal of dentistry*, vol. 77, pp. 106–111, 2018.

[15] T. Nakamoto, A. Taguchi, M. Ohtsuka, Y. Suei, M. Fujita, M. Tsuda, M. Sanada, Y. Kudo, A. Asano, and K. Tanimoto, "A computer-aided diagnosis system to screen for osteoporosis using dental panoramic radiographs," *Dentomaxillofacial Radiology*, vol. 37, no. 5, pp. 274–281, 2008.

[16] P. Chu, C. Bo, X. Liang, J. Yang, V. Megalooikonomou, F. Yang, B. Huang, X. Li, and H. Ling, "Using octuplet siamese network for osteoporosis analysis on dental panoramic radiographs," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2018, pp. 2579–2582.

[17] J. Lee, S. Adhikari, L. Liu, H. Jeong, H. Kim, and S. Yoon, "Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study," *Dentomaxillofacial Radiology*, vol. 48, no. 1, 2019.

[18] A. K. Jain and H. Chen, "Matching of dental x-ray images for human identification," *Pattern recognition*, vol. 37, no. 7, pp. 1519–1532, 2004.

[19] J. Zhou and M. Abdel, "A content-based system for human identification based on bitewing dental x-ray images," *Pattern Recognition*, vol. 38, no. 11, pp. 2132–2142, 2005.

[20] L. Lin, H. Lai, and W. Huang, "Dental biometrics: Human identification based on teeth and dental works in bitewing radiographs," *Pattern Recognition*, vol. 45, no. 3, pp. 934–946, 2012.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

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

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[26] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[27] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4974–4983.

[28] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[32] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.